THERE IS NO LOGIC THAT CAN BE SUPERIMPOSED ON THE CITY; PEOPLE MAKE IT, AND IT IS TO THEM, NOT BUILDINGS, THAT WE MUST FIT OUR PLANS.

JANE JACOBS, *THE DEATH AND LIFE OF GREAT AMERICAN CITIES*

INFORMATION IS THE OIL OF THE 21ST CENTURY, AND ANALYTICS IS THE COMBUSTION ENGINE

PETER SONDERGAARD, *SVP, GARTNER*

ERRORS USING INADEQUATE DATA ARE MUCH LESS THAN THOSE USING NO DATA AT ALL

CHARLES BABBAGE *THE EDUCATION INDUSTRY*

BALAMURUGAN SOUNDARARAJ

ESTIMATING FOOTFALL FROM PASSIVE WI-FI SIGNALS

CASE STUDY WITH SMART STREET SENSOR PROJECT

DOCTOR OF PHILOSOPHY

UNIVERSITY COLLEGE LONDON - UCL

DOCTOR OF PHILOSOPHY

DEPARTMENT OF GEOGRAPHY, UCL

I, *Balamurugan Soundararaj* confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

.....

Submitted on, November 2019

Dedicated to my parents,
S. Kalavathy and K. Soundararaj.

Contents

List of Figures

List of Tables

Glossary

- **Active vs Passive Collection** - Active collection is where the data collection process involves the active participation from the study subjects. In passive data collection, no such participation is required. For example, a web form based survey is an active data collection process while a script collecting data on user's data on a website is passive. This shall not be confused with primary and secondary data where the difference is mainly due to who collects the data from the source.

- **Anonymisation and Pseudonymisation** - The act of removing personal or identifiable information from the data. For example, removing the names and date of birth of people in a dataset. Anonymisation could be carried out in various ways two most popular approaches are generalisation and perturbation. Pseudonymisation is similar but the personally identifiable data is substituted with artificial identifier. The difference between them is that in anonymisation the personal information is permanently purged and there is no way of getting the information back. De-anonymisation is the reverse process of getting personal identifiable data from anonymised data.

- **Big Data** - Generally defined as data which could not be handled with regularly used tools and techniques. There are more nuanced definitions of big data depending on the dimension, context and industry. These concepts are discussed in detail in Section 4.1.

- **Bluetooth** - Short wavelength, low energy, short range wireless technology used for transferring data between electronic devices. They are generally used by mobile devices to connect to peripherals.

- **CDRC** - The Consumer Data Research Center is an investment from Economic Social Research Center, UK for working with consumer-related organisations to open up their data resources to trusted researchers thus enabling them to carry out important social and eco-

nomic research.

- **Cellular/Mobile Network** - Terrestrial, long range and wireless network which provides connectivity to mobile devices embedded in them. Most commonly used to denote the networks that provides telephony and internet services to mobile devices using radio waves.
- **Cryptographic Hashing** - This is the process of transforming a variable set of characters or contents of a file into a fixed length string (checksum). The process is usually one way and is not reversible without a rainbow attack using a database of checksums of all possible values. This is generally used for storing user passwords and to verify the integrity/ authenticity of data. In this thesis hashing functions are used for the purpose of anonymisaiton
- **Data Partner** - Local Data Company - the organisation which developed the Smart Street Sensor project in conjunction with CDRC. The organisation is solely responsible for the design, manufacture, installation and maintenance of the Smart Street Sensors.
- **Device Fingerprinting** - This is the process of identifying devices through forensic analysis. This commonly used to identify users from data collected by operational websites. This commonly done through identifying unique configurations of the devices such as screen sizes, versions of software installed, etc. In this thesis fingerprinting is used to denote such processes where unique devices are identified from randomised data.
- **Encryption** - This is the process of converting a information into unintelligible format using an algorithm (cipher) to prevent unauthorised access. The process is two way since the resulting cipher-text can be decrypted to restore information. The most common methods used is a public-key based encryption scheme known as 'Diffie Hellman key exchange'.
- **Footfall** - Number of people at a given location at a give time. In this thesis the term is used synonymous to footfall at a high street - which is only the pedestrians walking along the particular sidewalk of the high street and does not include people on the carriage way in vehicles.
- **High Street** - The primary business street in a town or a local area where most shops and commercial activity are located. This term is

often used to contrast and distinguish from 'Shopping centers' which are large designated areas earmarked for retail activity exclusively.

- **Localisation** - Though localisation could mean both GSM localisation and indoor positioning where the location of a device is indirectly derived from other information, in this thesis localisation is used mainly in the context of indoor positioning of mobile devices using techniques other than GPS. The localisation of objects is often relative to each other or to an object with an established position.

- **Location, Sensor and Install** - In this thesis and the Smart Street Sensor project, 'location' refers to a physical or geographic location in United Kingdom, 'sensor' denotes the electronic equipment used to collect data and 'install' refers to the time when a particular sensor was operating from a particular location.

- **Mobile Device** - A portable computing device such as a smartphone or tablet computer. This also includes wearables and other devices which have computing hardware and can operate independently of another device.

- **Personally Identifiable Information** - Any data that could potentially identify a specific individual. Any information that can be used to distinguish one person from another and can be used for de-anonymizing anonymous data can be considered personally identifiable information. In this thesis, the Media Access Control address is considered as sensitive personally identifiable information.

- **Positioning** - Positioning is the measurement of the absolute position (coordinates) of an object with reference to the earth. This is usually achieved through the use of Global Positioning System or similar systems.

- **Probe Request** - This is a special signal broadcast by Wi-Fi enabled mobile devices to elicit a response (probe response) from Wi-Fi access points which can receive them. The primary purpose of the probe request is to enumerate the available Wi-Fi networks and there can also be secondary purposes such as indoor positioning.

- **Real-time** - This is highly subjective concept which could be defined as the phenomena which occurs sufficiently immediately. In this thesis real-time is used to describe data which is collected within an hour after the event has occurred.

- **Sensor Configuration** - The way the sensor is installed at a particular location. This includes the this includes the position of the sensor in terms of height and depth, the material of shopfront etc.
- **Signal, Noise** - 'Signal' is used to denote the data received from devices which are within study area and 'Noise' to denote the data from devices outside the study area.
- **Smart Street Sensor Project** - This project is a comprehensive study of live footfall patterns across Great Britain using 1,000 Wi-Fi based sensors located in high streets across 81 towns and cities across the country. Smart Street Sensor is a Raspberry Pi based sensor designed and manufactured by the Data partner which collects data for the project.
- **Wi-Fi** - A family of medium range radio technologies based on the IEEE 802.11 family of standards which are generally used for wireless local area networking between devices.

Outputs From The Research

Publications

Book Chapter - Murcio, R., Soundararaj, B., & Lugomer, K. (2018). Movements in Cities: Footfall and its Spatio-Temporal Distribution. In Longley P., Cheshire J., & Singleton A. (Authors), *Consumer Data Research* (pp. 84-95). London: UCL Press.

Journal Paper - Soundararaj, B., Cheshire, J., & Longley, P. (2019) Estimating real-time high-street footfall from Wi-Fi probe requests, *International Journal of Geographical Information Science*.

Conference Paper - Soundararaj, B., Cheshire, J., & Longley, P. (2019). Medium Data Toolkit - A Case study on Smart Street Sensor Project. *In Proceedings of GIS Research UK*, New Castle, United Kingdom.

Conference Paper - Lugomer, K., Soundararaj, B., Murcio, R., Cheshire, J., & Longley, P. (2017). Understanding sources of measurement error in the Wi-Fi sensor data in the Smart City. *In Proceedings of GIS Research UK*, Manchester, United Kingdom.

Conference Paper - Soundararaj, B., & Zhu, D. (2019). Estimating pedestrian flow from footfall counts using Geo-propagation. *In Proceedings of Conference on Complex Systems*, Singapore[1].

[1] *Under consideration*

Conference Paper - Murcio, R., Soundararaj, B., & Lugomer, K. (2018). Trends in urban flows: An information theory approach. *In Proceedings of Conference on Complex Systems*, Cancun, Mexico.

Conference Paper - Soundararaj, B., Murcio, R., & Lugomer, K. (2016). Smart Street Sensor Footfall Project. *In Proceedings of Conference on Complex Systems*, Amsterdam, Netherlands.

Conference Paper - Soundararaj, B., & Lugomer, K. (2016). Smart Street Sensor Footfall Project. *In Proceedings of Oxford Retail Futures Conference*, Oxford, United Kingdom.

Engagement Activities

Event Talk - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Geo+data*, London, United Kingdom.

Event Talk - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Data Natives*, London, United Kingdom.

Conference Workshop - Estimating real-time high street footfall from Wi-Fi probe requests. (2018). *Smart Urban Policy Futures Workshop*, London, United Kingdom.

Exhibition - Real-time footfall in Bloomsbury. (2017). *big data Here - big data Week 2016*, London, United Kingdom.

Products

Research Dataset - Smart Street Sensor footfall data, *Consumer Data Research Center*. URL: https://data.cdrc.ac.uk.

Software - Soundararaj, B. (2018). Clicker - an Android app for manually counting pedestrian footfalls with precision and accuracy. *Google Play Store*. URL: https://play.google.com/store/apps?id=com.bala.manualcount.

Awards

Best Paper - Early Career Research paper at *GIS Research UK conference, New Castle* (2019).

Bursary - Yusuf Ali travel bursary (2017) *University College London*.

Prologue

Abstract

Measuring the distribution and dynamics of the population at granular level both spatially and temporally is crucial for understanding the structure and function of the built environment. In this era of big data, there have been numerous attempts to undertake this using the preponderance of unstructured, passive and incidental digital data which are generated from day-to-day human activities. In attempts to collect, analyse and link these widely available datasets at a massive scale, it is easy to put the privacy of the study subjects at risk.

This research looks at one such data source - Wi-Fi probe requests generated by mobile devices - in detail, and processes it into granular, long-term information on number of people on the retail high streets of the United Kingdom (UK). Though this is not the first study to use this data source, the thesis specifically targets and tackles the uncertainties introduced in recent years by the implementation of features designed to protect the privacy of the users of Wi-Fi enabled mobile devices. This research starts with the design and implementation of multiple experiments to examine Wi-Fi probe requests in detail, then later describes the development of a data collection methodology to collect multiple sets of probe requests at locations across London. The thesis also details the uses of these datasets, along with the massive dataset generated by the *'Smart Street Sensor'* project, to devise novel data cleaning and processing methodologies which result in the generation of a high quality dataset which describes the volume of people on UK retail high streets with a granularity of 5 minute intervals since August 2015 across 1000 locations (approx.) in 115 towns.

This thesis also describes the compilation of a bespoke *'Medium data toolkit'* for processing Wi-Fi probe requests (or indeed any other data with a similar size and complexity). Finally, the thesis demonstrates the value and possible applications of such footfall information through a

series of case studies. By successfully avoiding the use of any personally identifiable information, the research undertaken for this thesis also demonstrates that it is feasible to prioritise the privacy of users while still deriving detailed and meaningful insights from the data generated by the users.

Impact Statement

We live in the age of data deluge where data are generated at a pace that far exceeds our capacity to digest and analyse them. Putting these amounts of data to use within the constraints of available resources and time, is one of the biggest challenges faced by researchers today. The primary impact of this research is in solving this issue. This research utilised one such dataset - Wi-Fi signals generated by millions of mobile phones all around the year and available to anyone with a Wi-Fi receiver - then cleaned and processed them into highly granular and longitudinal information on the volume of footfall at retail high streets across the UK.

In converting the unstructured data into useful information, the research undertaken for this thesis developed two novel methods - one for filtering Wi-Fi signals based on their strength, and the other for grouping them based on their source mobile device. Moreover, this was achieved without revealing the identity of the users. These techniques enable researchers to deal with datasets exhibiting similar challenges such as Bluetooth signals, or records of people's clicking as they navigate through websites, etc. These methodologies and their results have been published in a peer reviewed journal *International Journal for Geographic Information Science* for the benefit of the wider community. They were also presented to the data partner who collaborated with the research unit - *Consumer Data Research Centre (CDRC)* - for the Smart Street Sensor project, and were considered for inclusion in the data partner's commercial project.

When dealing with the large and complex Wi-Fi dataset, the research designed and implemented a bespoke toolkit and a data processing pipeline comprising of open-source and free software which could be used by other researchers for use with similar datasets. The work on this 'Medium-data toolkit' was presented at the conference *Geographic Information Science Research UK*. Moreover the research directly led to the creation and maintenance of the *aggregated footfall* data product disseminated by CDRC[2], and has served as the data source for multiple research projects within and outside CDRC and UCL.

[2] *Local Data Company & UCL Smart Street Sensor Footfall Data: Research Aggregated data* - https://bit.ly/2FNGmo0

Apart from the technical impact, the primary output of the research - footfall volumes on retail locations - has commercial and policy impact for all the stakeholders involved with the retail industry in the UK. From this information comes a variety of insights: retailers can derive insights on the patterns of customer movement around their shops; landlords can find a reliable way to value their properties; local authorities gain a way to quantify and track the vibrancy of their retail centres over long periods of time; and consumers get information on which areas might be crowded at any given time. Finally, in the past 3 years, the outputs from this research have been disseminated to the broader academic community and industry through a series of paper presentations at conferences such as *GIS Research UK* and *Conference of Complex Systems*, talks at *Data natives*, *Geo+Data London*, and *Smart Urban Policy Futures Workshop*, industry events such as *Oxford Retail Futures Conference*, and public engagement events such as the *big data Here* exhibition.

Acknowledgements

I would like to thank my supervisors James Cheshire and Paul Longley for their guidance, support and relentless motivation. I am deeply grateful to my family who have supported me through my academic journey to whom I owe everything I have ever achieved. A very special gratitude goes to Anshita whose unwavering support and understanding got me through the most stressful times of my research. My sincere thanks to all my friends and colleagues from *Consumer Data Research Center* and *Department of Geography, UCL* who have helped me in the research. Most of the data collection carried out in this research were made possible by the help and support from our data partner - *Local Data Company*. I would like to thank *Economic and Social Research Council* for funding the research under the awards - ES/L011840/1 and 1625064. Last but not the least, I would like to thank the contributors to all the open source and free projects I have used in this research (appendix 7.6). I am forever indebted to them and hope to repay them by contributing back to the community for the rest of my career.

1

Introduction

Our understanding of the form and function of cities and the built environment has evolved significantly since the early twentieth century. What started as a field of research focused on the physical form of spaces and places, later moved towards modelling them as a function of the population that lives in them. Rather than viewing the built environment as infrastructure which need to be built, maintained and managed independently, cities have increasingly been viewed as the manifestation of the distribution and dynamics of the population embedded within them. The field was further broadened in the later part of the twentieth century to include the economic and social activity which happens within the fabric of the built environment. Moreover, with the dawn of the information age around the turn of the millennium, the built environment can now be viewed as the tangible result of information exchange; where cities can be seen as high density clusters of information exchange, in addition to as places with a concentration of physical infrastructure such as buildings and roads. This information revolution has not only changed researchers' understanding of the underlying forces of the built environment, but has also changed how they approach the task of measuring, analysing, modelling and managing it. The information revolution has provided researchers with numerous new technologies, methodologies, and tools. Perhaps most importantly however, is the unprecedented availability of comprehensive, granular data generated from fundamental functions of the built environment, such as human mobility, social interaction and economic activity. Availability of these data and tools has turned numerous disciplines upside down resulting in research which tackles problems using a bottom-up, 'data first' approach, rather than a more traditional top-down 'systems' approach.

We are currently in an age of 'data deluge' where the amount of data generated in the world far exceeds our capacity to analyse and

derive insights from them. This deluge of data has accelerated to such an extent that 90% of all the data ever generated in the world has been created in the last 2 years [1]. With the popularisation of wearable technologies and the 'internet of things', this trend is not expected to change any time soon. Moreover, many day-to-day activities of people such as banking, bill payments, public transport ticketing, taxi hire, social communications, and fitness tracking have been digitised and are generating large amounts of unstructured data as a consequence. As such, collecting data for some types of quantitative research has changed from a highly structured, designed endeavour to a low cost, scraping activity from data repositories heretofore relatively unused in terms of research beyond the purpose for which the data were initially collected. Most of the data collection activity has also become 'passive', i.e. collected without any effort from the participants. This has vastly increased the capacity of the data collection process, which has led to the emergence of 'big data' and consequentially, to the need for advanced and automated data-mining techniques to extract value from these vast datasets. The above two phenomena – the attempt to model the physical environment as a function of information exchange, and the unprecedented availability of data - has led to a significant volume of research wherein various data sources have been utilised to understand a variety of aspects of the built environment. For example, functional regions of a country has been derived from call detail record data [2], and population and demography have been studied through social media data such as that derived from Twitter.

[1] Ralph Jacobson. 2.5 quintillion bytes of data created every day. how does cpg & retail manage it?, Oct 2016. URL https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion

[2] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLOS ONE*, 5(12):1–6, 12 2010. DOI: 10.1371/journal.pone.0014248. URL https://doi.org/10.1371/journal.pone.0014248

1.1 *Challenges*

This frenzy of data generation and use is not without pitfalls. One of the major disadvantages in the attempts to repurpose the data is the risk to the privacy of the users whose data is being collected and analysed. With personal mobile devices becoming mainstream, almost every data point generated has a person behind it. The rush into the information age and the use of social media platforms has happened at a much faster rate than understandings of the ramifications to privacy of the participants could be properly understood. Even when the data collected does not contain personal information, most datasets can reveal personal and potentially sensitive information when linked with other sources of data. For example, anonymised cycle ride trajectories might not be interesting

information on their own, but when combined with other datasets such as taxi trips and payment information, the data can disclose the identity and residences of the people the data is about. This has prompted major concerns and backlashes from users and regulators in the past decade. These concerns are addressed in industry as well as research using both technology and regulation. From the technology perspective, all the stakeholders who generate, collect, or use the data try to use cryptography to anonymise, obscure, or encrypt any personal information as much as possible. In terms of regulation, legislation efforts such as the General Data Protection Regulation (GDPR) have been introduced to influence the behaviour of these stakeholders by introducing comprehensive rules and punitive measures for non-compliance. Though both these approaches ultimately try to protect user's privacy and personal information, they also pose one of the greatest challenges to research which uses passively collected user data. In the next 5 years, it can be expected that every freely available data source will be protected from the unfettered use which we see today. Wherever this protection is not possible, it can be expected that the data would be obscured or anonymised in order to remove any risk to the privacy of users, thus making it imperative that researchers adapt to these changes by looking for ways to overcome the challenges posed by them.

In addition to privacy concerns, this deluge of data introduces significant technological challenges as well. Both academia and industry have produced extensive 'big data' research which develops the theory, methods and tools to tackle the challenges posed by such large assemblages of data, in order to derive meaningful insights from them. This 'big data' research promises to solve a lot of the technological and logistical challenges incurred in many disciplines, but not without significant additional overheads in terms of cost and resources. In the case of research projects, blindly adopting the 'big data' methods without consideration, has the potential to cause more problems than advantages. The discipline of Geography, especially geographic information systems and science, has a long tradition of dealing with large datasets from the inception of the field, and the recent deluge of data has caused issues due to the complexity, latency and lack of structure of these new datasets, rather than their sheer volume. Hence, it is extremely important to be mindful while adopting the contributions from 'big data' discourse for research so that the solutions are implemented where the actual problems are located. There needs to be careful consideration when choosing or designing the

methods, tools and frameworks which are used to address the unique requirements of the new data sources. Moreover, there needs to be an inquiry into a framework for how these considerations are identified and addressed.

1.2 *Research Question & Methodology*

The motivation for the research began with the collection of the Wi-Fi probe requests at a national level within the 'Smart Street Sensors' (SSS) project. The primary objective of the project was to develop a business venture providing quantitative data on footfall to retailers across the country. This research was initially designed to supplement the above project by exploring the possibility of validating the data collection methodology and understanding the uncertainties and biases present in them. It was also designed to explore methods and analyses which provide insights and value to the retailers from the base footfall volumes. As discussed in the previous section, the preliminary analysis of the data collected revealed two major problems areas. First is the need to improve the accuracy of the footfall estimation by circumventing the MAC randomisation process and the second is the need to protect the privacy of user by developing methods that neither reverse engineer nor collect any personal data. With these two challenges in mind, the primary question posed for the research is as follows,

> "Can dynamics of footfall inferred from passively collected big dataset without putting the privacy of users at risk?"

In this context, this thesis works on answering the question while exploiting the opportunities presented above in the following ways: by first describing the collection of large volumes of passively generated data, then by solving the uncertainties in the data which arises due to their high variability and the mechanisms designed to protect the privacy of the users, and finally, by analysing the data to produce useful information regarding the distribution and dynamics of footfall in the country.

Contrary to regular methodology, where the research starts from the question, moves to a literature search, data collection, analysis and finally discussion and conclusions, this research starts from the availability of large comprehensive national level dataset. This research starts from this dataset, studies both the data and literature surrounding it in detail, finds gaps, problems and unanswered questions in the field, the tries

to solve or answer them. In this pursuit of understanding the nature of the data that were available without using personal data of users, the research also devises and conducts series of controlled experiments which provides valuable insight into improving the method that could be used for improving the estimation of people or footfall around the sensors. The effectiveness of the methodology was also tested using various sets of manually collected data on footfall information at sample locations.

1.3 Outline

The thesis starts with a broad and systematic literature survey on the topic of 'distribution and dynamics of human activity' in Chapter 2. In this chapter, major themes of research and their evolution in the past 30 years are identified along with the development of technologies which were employed. The literature review resulted in the identification of the best possible data source for further research, along with opportunities available for further research.

Having identified Wi-Fi as one of the most promising technologies for research, Chapter 3 explores Wi-Fi specification in detail, especially the 'probe request' mechanism. In addition to studying the standards and specification used to identify relevant data, the chapter also discusses the design and implementation of a series of small experiments to capture and analyse data in the real-world. Three sets of initial experiments were conducted and results from the experiments were used to conduct a longer and broader 'pilot study' which collected data from locations across London. The chapter also introduces the 'Smart Street Sensors' project - a national project which collects Wi-Fi data at a large number of retail locations. The chapter concludes with a detailed evaluation of all the data collected from the experiments and the Smart Street Sensor project, in terms of the bias, noise and uncertainties present in them.

Chapter 4 deals with processing the Wi-Fi data to remove the identified uncertainties in order to produce 'clean' and continuous information on the volume of footfall at the corresponding locations. The emphasis on not using personal data, or methods which can potentially reveal personal information, is firmly held throughout the chapter. In section 4.1, a framework for evaluating the 'bigness' of the data is discussed, and a 'data toolkit' for processing them is subsequently devised. In section 4.2, methods to clean the data into a realistic estimate of footfall

are discussed. In section 4.3, both the 'data toolkit' and methods are combined to architect a 'data pipeline' which digests the continuous stream of data from the SSS project into meaningful footfall numbers efficiently.

Chapter 5 details a variety of applications of the research across four major themes: an index for footfall across United Kingdom, the detection of events using changes in the volume of footfall, an estimation of the flow of pedestrians between locations derived from the changes in footfall volumes, and the identification of the nature and relationship between places along with possibilities for further research.

1.4 Impacts & Applications

The potential of creating such detailed, long-term, national-level footfall data as produced by this research is immense. Such information can be one of the major components in building a 'smart city', where the availability of detailed, real-time data on the built environment and its use is vital. It can also help us in our pursuit to accomplish a real-time census of people and their movement in the city. It can not only provide us with snapshots of the state of retail areas, but also help in measuring, modelling and manipulating them in real-time as a dynamic system which respond to interventions. We can even link these footfall data to other sources of data such as commercial consumer datasets and public transport statistics, in order to build a comprehensive picture on the health and efficiency of city-wide systems. Availability of such datasets can revolutionise academic research in fields such as urban planning, public policy and urban management, whereby the effect of interventions could be objectively measured and analysed.

Although this research did not try to explore the applications of this footfall data in detail, it hopes to serve as a solid basis for further studies in a various academic disciplines such as geography, business management, risk management, spatial analysis and computer science, which can employ the data to either derive insights about locations and context, or use the data as a reference/training source for validating methods and tools. Availability of such national level data on footfall volumes spanning continuously over years can also have a substantial impact on industries such as retail, transportation, real estate and information technology. As this research has a significant bias towards retail locations, the outputs can especially be of immense value for various stakeholders in

the retail industry such as *Retailers* who can get detailed information on when and where their customers shop which can lead to more efficient business operations, *Customers* who can be informed on the popularity of places and when to visit them, *Landlords* who can achieve a way to objectively evaluate their properties' values based on their location and also time, and *Local Authorities* who can be enabled to monitor and manage the health of their retail areas over longer periods of time.

2

Review of Literature

Understanding the scale, nature and dynamics of distribution of the population across space and time has been the central premise of academic research in various fields of study such as human geography, sociology, transportation, urban planning and managements. This granular knowledge of where people are and how they move is also critical in practical decision making in various industries such as real estate for valuing places, retail for business planning and risk management during emergencies for evacuation. A key challenge faced by these areas of research concerning the population at this scale is the collection of precise and accurate data in a timely manner. Though large structured datasets such as national census provides comprehensive coverage they are sparse temporally and understanding dynamics of population withing shorter periods is not possible. Alternatively, smaller datasets such as sample surveys and traffic counts are collected more frequently they are not comprehensive enough. This pursuit for identifying a data source which has the best features of both type of datasets started as an inquiry into methods to estimate and interpolate finer data from existing regional level aggregate data. As technology improved through the later half of twentieth century, research methodologies adopted the new tools and technologies to not only improve the quality of estimations but also to collect data with high granularity. Though new technologies provided immense opportunity in collecting large amounts of data which were previously impossible, they also introduced their own share of uncertainties. Hence it becomes imperative that the evolution of these techniques and methodologies are understood along with the research that employed them so that a rationale is built behind any further research. Moreover with the proliferation of mobile devices and wireless internet connectivity, every day to day activity is being digitised leading to the creation of large volumes of easily accessible data which are generated passively

in an unstructured manner. The users' acceptance to the collection and analysis of such data has also been improving until recently [1], but there has also been rising concerns regarding user privacy along with the development of more accurate methods to track behaviour. In this context, it is critical for research to balance these two: collecting relevant data and protecting user privacy, by choosing the right technologies and devising the appropriate methods.

[1] Alfred Kobsa. User acceptance of footfall analytics with aggregated and anonymized mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8647 LNCS, pages 168–179, 2014. ISBN 9783319097695. DOI: 10.1007/978-3-319-09770-1_15

In this chapter a systematic survey of literature in the broad and growing area of research concerned with quantifying the 'distribution and dynamics of human activity' has been carried out. The aim of this survey was to evaluate the stage at which the research is currently at, understand its evolution and progress through time and identify the possibilities that exists for future research. A comprehensive survey of over 300 publications which discuss this area of research was undertaken covering the major themes and trends in the last 40 years. These themes were discussed in detail to outline their contributions in the corresponding fields of study highlighting the opportunities and gaps in research that still exist. The timeline of publication of these research has also been studied to discuss the evolution of the research along with the changes in the technology landscape. These studies were then classified in terms of the major technologies employed by them to uncover the trends in how various technologies have been adopted and phasing out during this period. The primary objective was to understand the advantages and disadvantages of these techniques and to develop a theoretical framework for understanding when and how to use them effectively to answer research questions. Finally the literature survey was summarised focusing on the major research gaps that still exist and interesting new areas of research that has emerged recently where more research is warranted. These areas of research were also critically evaluated in terms of priority and feasibility leading to the development of questions and plan for this research thesis.

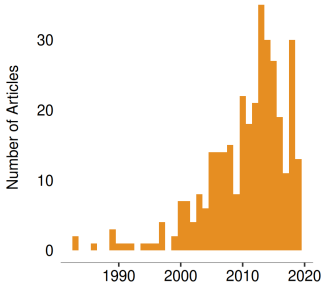


Figure 2.1: Growth of research in the area of 'Understanding distribution and dynamics of human activity' since 1980

Measured in the number of papers published

The set of works which discuss the use of mobile devices based technologies for studying topics in disciplines such as geography Li et al. [2016], urban analysis [Ratti et al., 2006], urban computing [Jiang et al., 2013] and other general applications and opportunities [Steenbruggen et al., 2013, Arribas-Bel, 2014], serve as our starting point for this literature survey. The search was then expanded from these reviews by navigating through their citation networks and identifying further research that are relevant. Though this did not provide a perfectly comprehensive set

of literature, it did provide a representative sample of all the different disciplines and directions of the research conducted in the area. Through this process, around 325 relevant research publications were identified which dealt with the collection, measurement, analysis, visualisation and discussion of population and their movement at a granular level. Research in this area started around 1950s where possibility of estimating day-time urban population at a granular level using existing broader data employing various estimation methods were discussed [2]. Though this served as a starting point, the pursuit of such granular data and their applications in corresponding fields didn't pick up until the start of the 21st century during the 'digital revolution' when personal computing become mainstream which was followed by the growth of internet. Figure 2.4 shows the yearly volume of research published since 1980 from which it can be observes that though there were some research conducted through 80s and 90s the real push forward came around beginning of the millennium when mobile phones adoption skyrocketed. In addition to the early 2000s, a substantial increase in interest can also be seen in the beginning of the next decade fuelled by the smartphone revolution which completely changed the research avenues in-terms of volume and types of data available and methodologies available to tackle them. While the mobile phone era put a device in every ones pockets, the smart phone era has armed them with immense data collection capabilities. The area of research is multidisciplinary encompassing academic interest and commercial applications in various disciplines and industries spanning across wide range of themes as discussed in section 2.1.

[2] Donald L Foley. Urban daytime population: a field for demographic-ecological analysis. *Social Forces*, pages 323–330, 1954. ISSN 0037-7732; and Robert C Schmitt. Estimating daytime populations. *Journal of the American Institute of Planners*, 22(2):83–85, 1956. ISSN 0002-8991

2.1 *Research Themes*

In this section we look at the major themes and questions tackled by this knowledge base. We start by classifying the research into the major and minor themes explored in them as shown in Figure 2.2. The treemap shows the volume of research in corresponding themes measured in terms of number of publications. We can observe that the research is conducted in five major areas - population studies focussing on the creating and utilising data on distribution and nature of human activity, mobility and interaction focussing on the changes in these distributions, understanding the nature and function of space from these distribution and change, methods and techniques which can be used to conduct the research and finally issues and solutions related to the privacy of the

users while conducting these research. We can also observe that most of the research apart from developing methods were conducted in the domain of human mobility and social interaction closely followed by the population distribution. In the following sections we discuss these in detail along with their sub themes with the following framework,

1. What are the major lines of questioning?
2. What has been done previously?
3. Where are the opportunities for further research?

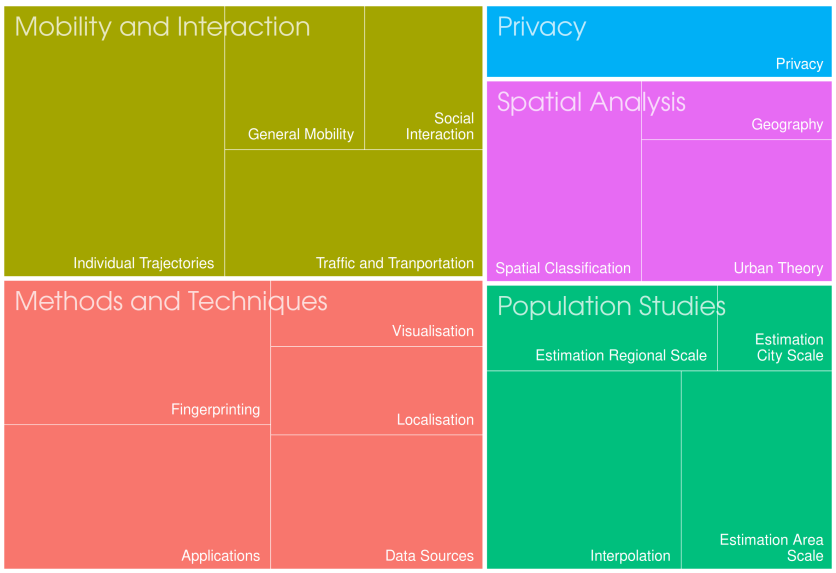


Figure 2.2: Tree-map showing the volume of research conducted under each major themes and their sub-themes.

Measured by the number of publications.

2.1.1 *Population Studies*

Though Foley [1954] and Schmitt [1956] started this line of research in 1950's with the discussion on estimating daytime population using broader datasets it was not until the 80s significant volume of research kicked off in this area of study. From 80s until mid 2000's numerous studies were conducted on measuring and studying the population at a granular level both spatially and temporally. The focus of the research around this time was primarily on interpolation from the larger datasets created using censuses, regional or national level sample surveys and other centrally collected sources of data. There have been numerous fairly successful attempts with methodologies where a broad dataset such as regional level population summaries and modelling or interpolating finer data from them by augmenting with other sources of data such as street networks [Reibel and Bufalino, 2005], remote sensing [Sutton, 1997, Yuan

et al., 1997, Chen, 2002] etc. Dobson et al. [2000, 2003], Bhaduri et al. [2002, 2007] and [Mennis, 2003, Mennis and Hultgren, 2006] are examples of such research methodology. These studies were almost done on a city scale or above with mostly modelling or interpolation methods since the data sources were few and were centrally collected.

Around 2005, there was a sharp shift in research where the interpolation methods were replaced by highly available granular data collected over cellular network. Studies were conducted on estimating population densities, presence of tourists, general activity pattens using data from cellular networks. Most of these research were conducted at a far larger geographic scale looking at things at an area level [Pulselli et al., 2008, Girardin et al., 2009, Phithakkitnukoon et al., 2010, Yuan and Raubal, 2016]. There were efforts in using device level sensors such as global positioning system(GPS), Wi-Fi and Bluetooth to detect population distribution and socio-geographic routines [Calabrese et al., 2010, Rose and Welsh, 2010, Farrahi and Gatica-Perez, 2010]. In terms of scale, there have been studies on looking at distribution of people at a highly granular level such as queue lengths [3] as well as broader level such as cities [4].

[3] Yan Wang, Jie Yang, Hongbo Liu, and Yingying Chen. Measuring human queues using wifi signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 235–237, 2013. ISBN 9781450319997. DOI: 10.1145/2500423.2504584. URL http://dl.acm.org/citation.cfm?doid=2500423.2504584{%}5Cnhttp://dl.acm.org/citation.cfm?id=2504584

[4] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45): 15888–15893, 2014. ISSN 0027-8424. DOI: 10.1073/pnas.1408439111

Around the 2015, along with the data collected directly from the mobile devices,the data that are generated by the users activity on these devices are became more important. Social media data such as twitter [Lansley and Longley, 2016b] and other consumer data such as loyalty cards [Lloyd and Cheshire, 2018], smart cards [Ordonez and Erath, 2012] etc. have also become a significant sources of data for such research. Recently, with increased concerns and legislation on privacy, there have been studies which go back to the effort of interpolating finer data from broader datasets but using more data and processor intensive technologies such as agent based modelling, deep learning, small area estimation [Crols and Malleson, 2019, Shibata and Yamamoto, 2019, Rao and Molina, 2015] etc.. Though there have been a lot of work done in most of the directions in this research area, the clear gap arises due to the absence of a continuous, granular and sufficiently longitudinal data-sets to complement the methodologies that have been developed.

2.1.2 *Human Mobility and Interaction*

Study of movement of people is one of the major areas of research which have significantly benefited from the decentralised collection of data at a granular level [5]. In addition to being useful in their own right, these data were in turn used to augment traditional models of travel

[5] M Castells. Grassrooting the space of flows. i wheeler, aoyama and warf [eds.] cities in the telecommunications age, 2000

behaviour, traffic and transport to provide a better understanding of human movement over time and space [Janssens, 2013]. The major themes of research within this area are, Movement of people in space and time with emphasis on understanding the built environment, social interaction between these people with a sociology perspective and traffic and transportation studies with a infrastructure perspective. There is significant volume of research which dealt with recording and analysing the trajectories of the users to understand their movement patterns enabled by the unprecedented availability of detailed data from mobile devices and this is discussed in detail along with the discussion of the technologies used in Section 2.3.

2.1.3 *Methodology and Techniques*

Research in this are focused around 5 major topics,

1. Localisation - Research into using the mobile devices and data generated from them as a cheaper alternative to Global Positioning Systems and remote sensing.
2. Data Sources - Identifying and formalising new data-sources as the technology develops
3. Application - Applying these identified data sources to answer questions and solve problems in different disciplines.
4. Visualisation - simplifying, visualising and interpreting these high volume of unstructured, noisy datasets.
5. Device fingerprinting - overcoming the difficulties posed by the anonymisation process and extract useful information.

Localisation of mobile devices without the use of expensive additional infrastructure such as GPS is one of the earliest ideas pursued in this aspect [Bulusu et al., 2000, He et al., 2003, Moore et al., 2004, LaMarca et al., 2005]. This research, when reversed, could also lead to the tracking of these devices in space without the aforementioned infrastructure thus providing a inexpensive, easy way to collect mobility data. The sensors which are already present in the phones such as Bluetooth [Bandara et al., 2004], Wi-Fi [Zarimpas et al., 2006], cellular radio [Dil and Havinga, 2011, Ahas and Mark, 2005] etc. have been considered to be used for localisation of the devices. This has been particularly important in the field of indoor localisation where GPS doesn't usually work [6]. When seen from the other perspective the same technologies and methods can enable us to collect presence and movement data on people indoors [Roy

[6] Nobuo Kawaguchi. Wifi location information system for both indoors and outdoors. In *International Work-Conference on Artificial Neural Networks*, pages 638–645. Springer, 2009

and Chowdhury, 2018a,b, Jia et al., 2019, Nikitin et al., 2019, Kulshrestha et al., 2019, Deng et al., 2018].

The identification of data sources started with looking at the 'real time' city examining the digital landscape created by the citizens their electronic devices [Townsend, 2000]. This was furthered by the notion of 'instrumenting' the city and developing methods and techniques under the umbrella of smart cities and internet of things [O'Neill et al., 2006, Sruthi, 2019]. Since there have been research looking at the wireless data collected from positioning technologies [Bensky, 2007] and cellular network [Kiukkonen et al., 2010, Steenbruggen et al., 2015] and even crowdsourcing as method of collection [Shin et al., 2013] leading towards a framework for computational urban planning [Kontokosta, 2015]. With the effort to formalise them as valid sources of data, there have also been research looking at the biases in them such as mobile phone ownership [Wesolowski et al., 2013, Kobus et al., 2013].

Identifying and fingerprinting unique devices and users from noisy, unstructured data is another area of active research under methodologies and techniques [7]. The majority of the work has been done as an extension of localisation where the GPS-less positioning leading to finger printing people and their movement out of the data [Pang et al., 2007a, Pappalardo et al., 2015]. Additionally there are work looking at the tracks collected from Wi-Fi or mobile data and extract unique users out of them [Girardin et al., 2008, Eagle and Pentland, 2009, Jiang et al., 2012]. It is also demonstrated that it is possible to wireless technologies can be used to detect even device free entities [Elgohary, 2013]. These localisation and clustering techniques can also be used for socio-geographical analysis and to understand the patterns of activity of people [Licoppe et al., 2008]. There has been a good deal of security research on the robustness of the anonymisation techniques while revealing methodologies to overcome limitations imposed by them [Mathieu Cunche, 2016, Chothia and Smirnov, 2010, Krumm, 2007]. Cheng and Wang [2016] was one of the first to look into devising a method to do this in a non-intrusive way which are further extended by Di Luzio et al. [2016], Adamsky et al. [2018] and Dai et al. [2019]. This is currently an active field of research and there is immense opportunity for further research.

Visualising the temporal dynamics of data collected on human activities through decentralised processes poses significant challenges when approached with traditional cartographic concepts [8]. Digital media especially animation has been explored as an option to solve for the temporal

[7] Bin Jiang and Xiaobai Yao. Location-based services and gis in perspective. *Computers, Environment and Urban Systems*, 30 (6):712–725, 2006. ISSN 0198-9715; and Lin Liao. *Location-based activity recognition*. PhD thesis, University of Washington, 2006

[8] Alan M MacEachren and Menno-Jan Kraak. Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28 (1):3–12, 2001. ISSN 1523-0406; and Elaine J Hallisey. Cartographic visualization: an assessment and epistemological review*. *The Professional Geographer*, 57(3): 350–364, 2005. ISSN 0033-0124

dimension [Morrison et al., 2000, Lobben, 2003] but is bound by the cognitive limits of the viewer [Harrower, 2007]. There have been approaches proposed around animations of generated surfaces [Kobayashi et al., 2011] and network-based visualizations [Ferrara et al., 2014] leaving gaps in research for new methods in dynamic geographic visualisation [Fabrikant, 2005] and visualising path and flow of phenomena [Thomas, 2005], particularly of the mobility data collected from cellphones [Sbodio et al., 2014]. This provides us with a promising opportunity for research in methods for visualising high frequency, hyper-local pedestrian data within the limits of cognition of the viewer.

2.1.4 *Spatial Analysis - Theory and Modelling*

Traditional and modern geography was dominated by the study of centrally collected data acquired through extensive field surveys and remote sensing. In the last two decades, a significant paradigm change has been introduced by the availability of unprecedented amount of data generated by unconventional sources such as mobile phones, social media posts etc. This move to the postmodern geography has been accompanied by a change in our understanding of the built environment and human geography from a static point of view to a more dynamic definition [9]. This definition is based on the bottom-up mechanisms which make human activity such as information exchange and economy to manifest in the physical built environments as argued by [Batty, 1990, 1997, Batty et al., 2012] and [Batty, 2013a,b].

[9] Edward Soja. Postmodern geographies, 1989. URL http://books.google.com/books?id=sNcRAQAAIAAJ

This transition into the digital age [Graham and Healey, 1999, Tranos and Nijkamp, 2012, Tranos, 2013] has changed the politics of space and time [Massey, 1992] and been more pronounced in the study of urban built environment where technology has redefined the concepts of place and space [Graham and Marvin, 2001, 2002, Sassen, 2001]. With the ability to collect and analyse of data on large complex systems in real-time [Graham, 1997], we are exploring the possibilities of understanding their structure and organisation using concepts of complexity theory [Bettencourt, 2013, Portugali et al., 2012] with more emphasis on their temporal patterns such as the argument towards finding the pulse of the city [Batty, 2010]. With the population getting more and more connected [Castells, 2010], the nature of space/place is being dynamically defined by the population themselves [Giuliano and Small, 1991] and vice versa [Zandvliet and Dijst, 2006]. This flood of hard data [10] was accompanied not only by optimism in its potential [Thomas, 2001] but

[10] Nature Editorial. A flood of hard data. *Nature*, 435:698, 2008. ISSN 0028-0836. DOI: 10.1038/453698a

also by the questions raised on the challenges in handling the diverse, large scale, non standardised data it produces and the usefulness or representativeness of the resulting analysis [Miller, 2010, Arribas-Bel and Sanz-Gracia, 2014]. However, availability of such data has impressive uses in urban studies [Bettencourt, 2014] especially with advancement of new technologies [Steenbruggen et al., 2013] and possibility of distributed, crowdsourced data collection [Lokanathan and Gunaratne, 2015].

2.1.5 *Privacy*

The ubiquity of personal devices and digitisation of day to day activities through these mobile devices [McMeel, 2018] has provided many opportunities for researchers and industry for collecting, analysing and deriving inputs from them. However at the same this also increased the risk of infringement on privacy of the users whose data is being collected [11]. There is immense value in uniquely identifying and profiling information on people for specialised purposes such as security [Cutter et al., 2006] and law enforcement [Dobson and Fisher, 2003] but also has extreme risks associated when not handled with care [VanWey et al., 2005]. Strictly protecting personal information while ensuring the information is usable for research by maintaining the uniqueness in the data is the major concern which was addressed by devising frameworks for secure practices in confidentially collecting and using the location data [Duckham and Kulik, 2006, Tang et al., 2006, Lane et al., 2014]. Some efforts sought to accomplish this task through cryptographic hashing algorithms [Pang et al., 2007b] while others aimed to thwart identification and tracking at the device level by techniques such as MAC randomisation [Gruteser and Grunwald, 2005, Greenstein et al., 2008]. Finally though getting consent of users for the collection and use of such information from their mobile devices is challenging, there is a significantly improved acceptance when the process offers value in return such as discounts and monetary benefits [Kobsa, 2014].

There is opportunity in this area for research in applying the cryptographic solutions along with the privacy preserving frameworks to arrive at methods which can extract useful information out of large personal data while obscuring or anonymising them.

[11] T Scott Saponas, Jonathan Lester, Carl Hartung, Sameer Agarwal, and Tadayoshi Kohno. Devices that tell on you: Privacy trends in consumer ubiquitous computing. In *Usenix Security*, volume 3, page 3, 2007; and John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009

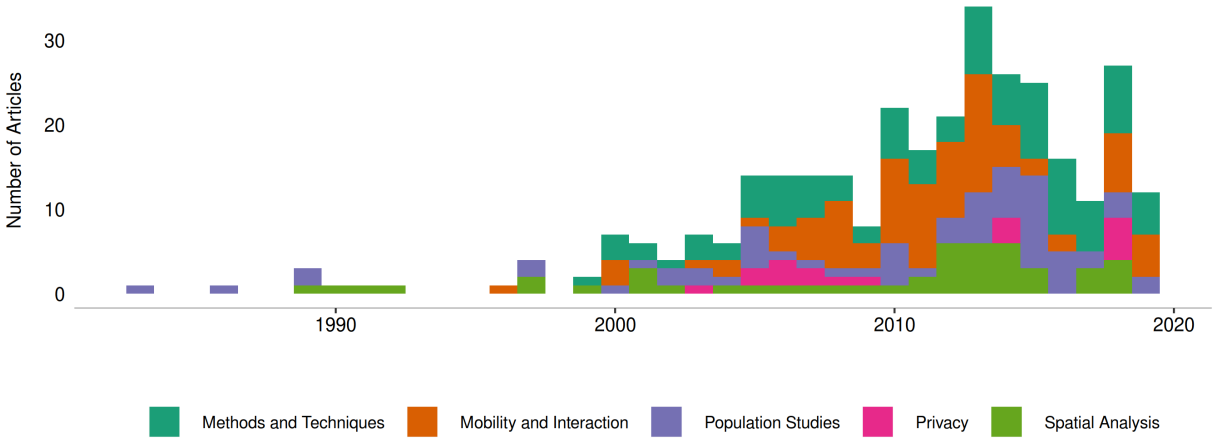


Figure 2.3: The evolution of research since 1980 categorised based on their major theme.

2.2 *Research Trends*

Figure 2.3 shows the volume of research done in this topic since 1980 categorised based on their major themes discussed earlier. We can observe that following distinct trends exist in the research, which evolved around the development of technology in the last two decades. Until 90s the research was mostly centered around population studies on estimating and interpolating granular spatial and temporal information from larger and cross sectional datasets such as census and sample surveys. The period between 2000-2010 there was interest in potential of the new data generated by the digital revolution. We can categorise this as the 'mobile era' where carrying mobile devices become mainstream. This explosion of research coincided with mobile phones becoming more popular and ubiquitous with population in urban areas and was around development of methods and techniques to utilise the data generated from them. There were also extensive studies in using the datasets to understand human mobility along with a rising concern in the privacy of the users who's data which are being used for these studies.

The release of iPhone in 2008 and the increase in the share of 'smartphones' in the next 10 years sparked the 'smartphone' era. The change made sure that all the mobile devices gaining numerous capabilities such as internet connectivity over Wi-Fi and mobile network, location awareness with global positioning system, movement recognition with accelerometers and connectivity other 'wearable' devices through Bluetooth. This also lead to the digitisation of lifestyle where every aspect of the life being done through these devices over internet while generating huge amount of data on these activities. This sparked the large volume

of research on the form and function of space by studying this data and on the dynamics of human population in space and time in the next 5 years.

These research were particularly centered around tracking the trajectory of people using the mobile devices they carry with them as the smartphones made it easier to collect the necessary data directly from them rather than depending on a centrally collected datasets from mobile carriers. With the theoretical limit to predictability in human mobility quantified by Song et al. [2010b], the focus on urban mobility has been declining in the past few years which has led to a renewed interest in population studies at a local-local level in real-time. In addition to using the data from the mobile devices, these studies have also been exploring the use of large assemblages of consumer data that are being generated in this connected mobile environment and linking them together to create a fuller picture [12]

[12] Paul Longley, James Cheshire, and Alex Singleton. *Consumer Data Research*. UCL Press, 2018

Finally, with the increase in use of personal data, there has also been an increase in research regarding the privacy of the users. Along with this, the mobile devices and subsequently the data generated by them are more and more anonymised so that the users cannot be tracked or identified at a personal level. This has given rise to the new trend in research to devise methods to overcome this anonymisation and at the same time research which considers these methods as vulnerabilities and find solutions to make the anonymisation process more robust. There is clear need for methods which anonymise the data sufficiently to protect the identity of the users and at the same time enable us to conduct research in measuring studying population distribution and movement at a granular level.

2.3 *Techniques and technology*

When we look at the literature from the technology perspective, we observe that over the years, the research continuously picks up and applies recent technological developments in the pursuit of understanding the distribution of human activity and population. Figure 2.4 shows the distribution of the research in terms of the main technique/ technology used over the past 40 years. We observe that the earliest attempts started from the exploration of using interpolation and modelling techniques on a broader dataset. As the need for more granular datasets increased there were attempts to devise and utilize bespoke solutions to generate them.

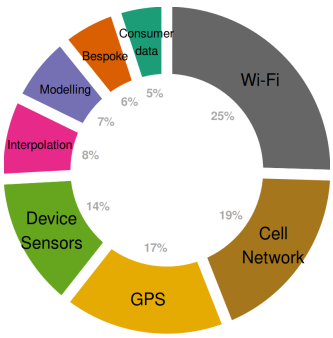


Figure 2.4: Distribution of research across various techniques and technologies
Measured in the number of papers published

When mobile devices became mainstream, the focus shifted to utilize the relevant components of the mobile infrastructure. A significant number of studies were done in utilising data collected from the mobile network, sensors in the mobile devices, especially GPS and Wi-Fi, in addition to the social media content generated from these devices. A detailed account of these studies is given below,

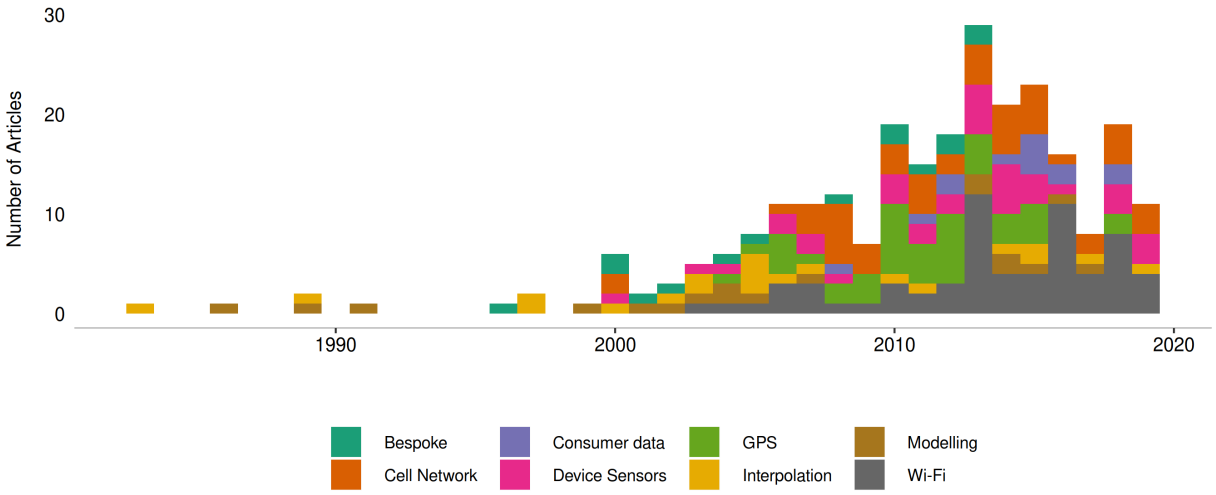


Figure 2.5: The evolution of research since 1980 in terms of the the technology used in the research.

2.3.1 *Interpolation and Modelling*

Attempts in using the existing data collected through traditional methods such as census and large scale sample surveys to create spatially and temporally granular and detailed estimates were carried out by applying various interpolation methods such as pycnophylactic, dasymetric interpolation [Tobler, 1979, Mennis, 2003, Mennis and Hultgren, 2006, Hawley and Moellering, 2005, Tapp, 2010, Wismans et al., 2017] along with spatial [Lam, 1983, Martin, 1989, Martin et al., 2015] and temporal interpolation techniques [Glickman, 1986]. These methods along with supplementary data such as remote sensing imagery [Sutton et al., 2001, Chen, 2002] and street networks [Reibel and Bufalino, 2005] were shown to be useful in producing detailed granular population maps at various scales with varying degree of success [Dobson et al., 2000, Bhaduri et al., 2002, Dobson and Fisher, 2003, Bhaduri et al., 2005, 2007]. These approaches have been employed in various applications such as econometric studies [McDonald, 1989], studies on public health [Hay et al., 2005], emergency management [Kwan and Lee, 2005] and flood risk estimations [Smith et al., 2016].

In addition to these interpolation techniques classic modelling tech-

niques can also be used to estimate daytime populations and demographic structure at hyper-local scales [Jochem et al., 2013, Jia et al., 2014], urban scales [Alahmadi et al., 2013, Abowd et al., 2004] and regional scales [Foley, 1954, Schmitt, 1956, Singleton and Longley, 2015, McCormack, 2017]. The granular data created with such modelling techniques are shown to be useful in urban planning and management [Parrott and Stutz, 1999], emergency management [Alexander, 2002, Cutter et al., 2006] and in modelling traffic and transportation [Lefebvre et al., 2013]. These interpolation and modelling techniques along with granular data produced are also used in classifying spatial areas and hence understanding the structure of cities in general [McMillen, 2001, 2004, Lee, 2007, Arribas-Bel, 2014]. Though being useful, these techniques are still shown to have limitations and uncertainties [Nagle et al., 2014], which mostly arise from the nature of the input data employed. This leads us to the need for more detailed and frequent collection of data.

2.3.2 *Bespoke technologies*

Following this need, there has been efforts to use bespoke or specialised technologies such as cameras [Cai and Aggarwal, 1996, Heikkilä and Silvén, 2004, Kröckel and Bodendorf, 2012], Lasers [Zhao and Shibasaki, 2005, Arras et al., 2008] and radio frequency receivers [Bahl and Padmanabhan, 2000, Yang et al., 2013, Chothia and Smirnov, 2010, Bulusu et al., 2000, Dil and Havinga, 2011] to measure human activity. But the major problem with such solutions is the cost and effort involved in designing and implementing them at urban and regional scales comprehensively. Moreover, being specialised and centralised they tend to be challenging to maintain and update as the technological landscape change. This gives us the need to identify and use techniques which are more general in nature and can be used for longer periods of time which are cheap to install to achieve a more comprehensive coverage.

2.3.3 *Cellular Network*

The rise of mobile phones as ubiquitous personal devices for the broader population has provided us with a viable alternative for collecting data with finer granularity at large scales. Mobile infrastructure consists of both the 'network part', built and managed by the service providers, and the 'user part', which is the phones owned by the users' themselves. The network part, in addition to providing connectivity to the users, also collects information on these devices actively such as communication

between the users and passively such as when the phones themselves move from tower to tower. The mobile devices themselves have a variety of sensors such as accelerometer to identify movement, compass to identify orientation, GPS receiver to deduct geographic position, etc. They also have various communication capabilities such as cellular, Wi-Fi, Bluetooth and Near field communications (NFC) etc. Both of these sensors and communication capabilities can be used as sources of data themselves. With the growth of mobile devices and the infrastructure surrounding it, there has been significant effort in utilising data generated by every component of this complex infrastructure.

The first set of research started to use the cellular network data for urban studies [Jiang et al., 2013, Steenbruggen et al., 2015, Lokanathan and Gunaratne, 2015, Calabrese et al., 2015, Reades et al., 2007]. Even though this approach has been acknowledged to have inherent biases such as ownership bias across particular demographic groups [Wesolowski et al., 2013] the relative advantages such as coverage made them excellent sources of data. Visual exploration of use of such data using interactive interfaces to evaluate quality of service and scenario testing has been tested for the optimisation of public transport [13]. Such network data with the active and passive information collected from them can be used to create trajectories of people [14], detect their daily routine [15] and classify those routes in terms of function [Becker et al., 2011a]. It was also demonstrated to be useful in understanding overall mobility and flow of people and information [Candia et al., 2008, Krings et al., 2009, Simini et al., 2012, Zhang et al., 2019]. These data can be used to identify asymmetry in flow of people spatially [Phithakkitnukoon and Ratti, 2011], estimate volume and pattern of road usage [Bolla et al., 2000, Wang et al., 2012] and by augmenting the topology to optimise operations [Puzis et al., 2013]. Such datasets have been extensively used in traffic and transportation research to derive origin-destination matrices [Caceres et al., 2007, Mellegard et al., 2011, Iqbal et al., 2014], travel time estimation [Janecek et al., 2012] and traffic status estimation [Demissie et al., 2013, Grauwin et al., 2015].

It has been shown that mobile network data can be used to uncover nature of the population such as tourists in specific areas [Girardin et al., 2008] and the interaction between the people in the study area [Campbell et al., 2008]. The structure [Onnela et al., 2007a,b], geography [Lambiotte et al., 2008] and dynamics [Hidalgo and Rodriguez-Sickert, 2008] of such networks have been studied and demonstrated to be useful in predicting their change [Wang et al., 2011, Vajakas et al., 2018] over

[13] Marco Luca Sbodio, Francesco Calabrese, Michele Berlingerio, Rahul Nair, and Fabio Pinelli. All aboard: visual exploration of cellphone mobility data to optimise public transport. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 335–340. ACM, 2014. ISBN 1450321844

[14] Johannes Schlaich, Thomas Otterstätter, and Markus Friedrich. Generating trajectories from mobile phone data. In *Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies*, 2010

[15] Andres Sevtsuk and Carlo Ratti. Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17 (1):41–60, 2010. ISSN 1063-0732. DOI: 10.1080/10630731003597322. URL http://www.tandfonline.com/doi/abs/10.1080/10630731003597322

time. This social networks and their spatio-temporal structure can also be used for classification of land use [Pei et al., 2014, Jia et al., 2018], assessment of spatial patterns [Reades et al., 2009, Steenbruggen et al., 2013] and understanding the broader spatial structure of cities [Louail et al., 2014, Arribas-Bel and Tranos, 2015] and regions [Arhipova et al., 2018]. The data collected from the cellular network when examined at granular levels such as inter-personal communication and economic activity can be used to create estimations of micro area-level population density [16] and also the characteristics [17] and the nature of the activity [Phithakkitnukoon et al., 2010]. Aggregated human activity measured from such research in turn can be used to measure and model population dynamics and land use density and mix at broader level [Jacobs-Crisioni et al., 2014, Tranos and Nijkamp, 2015, Tranos et al., 2018]. The spatial patterns thus uncovered can then be applied to urban planning [Becker et al., 2011b] whilst the temporal patterns uncovered have immense utility for the disciplines such as epidemiology. For example, population influxes measured from changes in mobile network usage can be used to model spread of diseases [18].

Though the mobile network provides much more granular and accurate data than interpolation techniques, it is not without its limitations [Yucel, 2017]. The distribution of network infrastructure usually follows the purposes of service coverage and follows commercial decisions. This introduces systematic biases in the data passively collected through them. Moreover, the data actively collected through them has bias based on the volume of usage of services by the customers which can vary widely spatially, temporally and also based on demography. In addition to this because of the coverage, the data collected from mobile service providers pose immense privacy risk when linked to other sources of consumer data. This makes collection of data directly from the devices using the sensors on the device much more robust in certain cases.

2.3.4 *Mobile Sensors*

The most prominent sensors and capabilities present in mobile devices that can be used for distributed urban sensing are Cellular radio, Bluetooth, Wi-Fi, GPS, accelerometer and compass [19]. Since cellular radio is managed by the cellular network and covered in mobile network data, we explore the research done with other sensors. In contrast to planned actively collected data, data passively collected via a distributed network of general purpose devices tends to be larger and more temporally dy-

[16] R Pulselli, P Ramono, Carlo Ratti, and E Tiezzi. Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *Int. J. of Design and Nature and Ecodynamics*, 3(2):121–134, 2008; and Yibin Ng, Yingchi Pei, and Yunye Jin. Footfall count estimation techniques using mobile data. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 307–314. IEEE, 2017

[17] Fabien Girardin, Andrea Vaccari, Alexandre Gerber, Assaf Biderman, and Carlo Ratti. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *Intl. Conference on Computers in Urban Planning and Urban Management*, 2009

[18] Caroline O Buckee, Andrew J Tatem, Justin Lessler, Ottar N Bjornstad, Bryan T Grenfell, Janeth Kombich, Nathan Eagle, C J E Metcalf, and Amy Wesolowski. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences*, 112 (35):11114–9, 2015. ISSN 0027-8424. DOI: 10.1073/pnas.1423542112. URL http://doi.org/10.1073/pnas.1423542112

[19] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010; and Enwei Zhu, Maham Khan, Philipp Kats, Shreya Santosh Bamne, and Stanislav Sobolevsky. Digital urban sensing: A multi-layered approach. *arXiv preprint arXiv:1809.01280*, 2018

namic. For example, an organised survey conducted every month to understand interpersonal communications between people in a team of 50 will result in a 2500 records a month. The same task is done through collecting data on email communication sent by them will result in a same volume records in a day. The challenges and solutions on collecting and analysing such large-scale longitudinal data are discussed by [Laurila et al., 2012, Antonic et al., 2013]. The real time nature of such data also gives us the opportunity to monitor and understand the city in much smaller temporal scales [Townsend, 2000, O'Neill et al., 2006] and the representativeness of such datasets have also been explored [Shin et al., 2013, Kobus et al., 2013]. Data generated from communication networks can be used to understand the structure of urban systems which are becoming increasingly border-less [20]. Similar to the network based data, it can help in understanding human mobility [Asgari et al., 2013, Amini et al., 2014, Zhang et al., 2014] through mining trajectory patterns [Giannotti et al., 2007] and socio geographic routines [Farrahi and Gatica-Perez, 2010]. It is also useful in various traffic and transportation applications for monitoring roads [Mohan et al., 2008] and estimating traffic [Cheng et al., 2006], uncovering regional characteristics [Chi et al., 2014] and extracting land use patterns [Shimosaka et al., 2014]. Apart from GPS and Wi-Fi, there have been efforts in exploring other possibilities such as Bluetooth for location [Bandara et al., 2004, Becker et al., 2019] and aggregate detected Bluetooth activity to monitor freeway status [Haghani et al., 2010]. There have also been successful implementations of frameworks to predict movement of people by combining Wi-Fi and Bluetooth [21]. But owing to shorter range and requirement of active engagement from the user where they have to actively start the device pairing process, Bluetooth is much less preferable for large-scale data collection than GPS or Wi-Fi. The research on GPS and Wi-Fi based studies are discussed in more detail below.

[20] Luca Bertolini and Martin Dijst. Mobility environments and network cities. *Journal of urban design*, 8(1):27–43, 2003

[21] Long Vu, Quang Do, and Klara Nahrstedt. Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*, pages 54–62. IEEE, 2011

2.3.5 *Global Positioning System*

In addition to providing a user's location to applications such as maps and navigation, the GPS capability in mobile devices in tandem with Wi-Fi can also maintain a continuous list of locations visited by the device over long periods of time. It works mostly in the background and requires almost no active input from the user to operate. Though very convenient for collecting data, due to the privacy risks associated with it, GPS is often one of the resources in a device that requires explicit

user permission to be accessed. The concepts and methodologies for collecting such data were set out by [Asakura and Hato, 2004] and there have been attempts to collect this rich data from volunteers at a large scale along with ancillary data [Kiukkonen et al., 2010] and provide a location based service application for the collection of data [Ratti et al., 2006, Jiang and Yao, 2006, Ahas and Mark, 2005].

The accuracy, convenience and being designed for navigation makes GPS one of the most used technologies for mobility studies [22]. It has been used to analyse and understand individual mobility patterns [Neuhaus, 2010, Shin et al., 2010], which have been shown to have a high order of regularity in spite of the complexity [Brockmann et al., 2006, Song et al., 2010a]. There have been efforts to use this regularity to predict the future location of people [Monreale et al., 2009, Calabrese et al., 2010]. The limitations of predictions have also been quantified [Song et al., 2010b]. There have been successful efforts in extracting behaviours and patterns from such trajectory data [Liu et al., 2010, Cho et al., 2011, Hoteit et al., 2013, Pappalardo et al., 2013] along to understand individual patterns from large assemblages [Giannotti et al., 2011, Calabrese et al., 2013] and vice versa [Wirz et al., 2012]. In traffic and transportation, GPS trajectory from mobile devices is used to estimate [Calabrese et al., 2011] and expand [Jing et al., 2011] origin-destination matrices, detect the mode of travel [Gong et al., 2012, Rossi et al., 2015] and calibrate existing spatial interaction models [Yue et al., 2012] . Since the data is collected at the device level and depends on the activity of the individual, it can be de-anonymised to reveal the nature of the owner of the devices. The possibilities of detecting the activity of the individual from trajectory information is demonstrated by [Liao, 2006, Krumm, 2007]. Patterns [Jiang et al., 2012] and structures in routines [Eagle and Pentland, 2009] can be extracted from these trajectories and can be used for socio geographic analysis of the population [Licoppe et al., 2008, Chen et al., 2018]. It can also utilised in classification of the population at a particular location at a given time [23]. Being inherently spatial and activity driven, GPS trajectories have been shown to be useful to identify [Bao et al., 2012], characterise [Wan and Lin, 2013] and automatically label [Do and Gatica-Perez, 2014] significant places of interest. It can also be used for land use detection [Toole et al., 2012, Zhang et al., 2018], classification [Jiang et al., 2015] and the study of urban morphology [Kang et al., 2012]. These GPS trajectories have been shown to be useful in estimating population dynamics at local level and within short durations during social events

[22] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008. ISSN 0028-0836. DOI: 10.1038/nature06958. URL http://www.nature.com/nature/journal/v453/n7196/full/nature06958.html{%}5Cnhttp://www.nature.com/nature/journal/v453/n7196/pdf/nature06958.pdf

[23] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6:8166, 2015. ISSN 2041-1723. DOI: 10.1038/ncomms9166. URL http://www.nature.com/ncomms/2015/150908/ncomms9166/full/ncomms9166.html{%}5Cnhttp://www.nature.com/doifinder/10.1038/ncomms9166

[Calabrese et al., 2010, Kim and MacEachren, 2014, Deville et al., 2014]. When combined with other data sources can be useful to understand relationship between spatial areas [Long and Thill, 2015].

From the literature we see that GPS is one of the most precise and accurate user side methods of collecting location of mobile devices. In addition, the data collected is well understood and collection methodologies can be scaled up with minimum resources. That being said, it is also well established that urban sensing methods using GPS of mobile devices has problems of enhanced risk of breach of privacy when executed passively and need explicit user engagement when executed actively.

2.3.6 *Wi-Fi*

Wi-Fi is a wireless network connection protocol standardised by IEEE [2016]. It is a distributed server-client based system where the client connects to access points (AP). Every mobile device in the network has a unique hardware specific MAC address, which is transmitted between the device and AP before the connection is made. The key feature of Wi-Fi infrastructure is that the network is distributed and the APs can be set up and operated by anyone locally unlike mobile networks. Since they are primarily used for Internet service provision, the protocol has priority for continuity of connectivity so the devices constantly scan for new and better connections. This is done through a probe request, which is detailed in later sections. With this background we can see that Wi-Fi provides a fair middle ground between an entirely network driven approach such as cellular network to an entirely user driven approach such as GPS. Since the network infrastructure is distributed and deployed for Internet it offers more coverage than most of the technologies discussed except or cellular network. It is also very resilient and can encapsulate and reinforce civic space in cities [24].

[24] Paul M. Torrens. Wi-fi geographies. *Annals of the Association of American Geographers*, 98(1):59–84, 2008. DOI: 10.1080/00045600701734133

Though Wi-Fi is a location less technology, there are reliable methods to trilaterate the location of the device by the signal strength and the locations of APs known through either targeted surveys or crowdsourced volunteer effort [He et al., 2003, Moore et al., 2004, LaMarca et al., 2005, Dinesh et al., 2017, Lin and Huang, 2018]. This can overcome the usual shortcoming of GPS, which struggles for precision and accuracy in indoor and densely built environments [Zarimpas et al., 2006, Kawaguchi, 2009, Xi et al., 2010]. Utilising this, we can easily and quickly estimate trajectories of the mobile devices just using the Wi-Fi communication the device has with multiple known APs [25]. This can be used similar to the

[25] Zhuliang Xu, Kumbesan Sandrasegaran, Xiaoying Kong, Xinning Zhu, Jingbin Zhao, Bin Hu, and Cheng-Chung Lin. Pedestrain monitoring system using Wi-Fi technology and rssi based localization. *International Journal of Wireless & Mobile Networks*, 5(4):17, 2013

GPS trajectories to understand individual travel patterns [Kim et al., 2006, Rekimoto et al., 2007, Sapiezynski et al., 2015], crowd behaviour [Abedi et al., 2013, Mowafi et al., 2013, Shu et al., 2017], vehicular [Lu et al., 2010] and pedestrian movement [Xu et al., 2013, Fukuzaki et al., 2014, Wang et al., 2016, Taylor et al., 2019]. It can also be used in transportation planning and management to estimate travel time [Musa and Eriksson, 2011, Håkegård et al., 2018] and real time traffic monitoring [Abbott-Jard et al., 2013].

Being a general network protocol designed to be used by mobile devices, Wi-Fi devices relay a range of public signals known as probe request frames on regular intervals throughout its operation, for the purpose of connecting and maintaining a reliable and secure connection for the mobile device [26]. These signals can be captured using inexpensive customised hardware, non-intrusively and in turn to be used for numerous applications. In addition to a uniquely identifiable MAC address, these signals include a range of other information which when combined with the temporal signatures of the signals received can help us understand the nature and identify the devices which are generating these signals. These device/user fingerprinting techniques are demonstrated by [Franklin et al., 2006] and [Pang et al., 2007b] and the unique MAC addresses and associated information can successfully track people across access points [27], their trajectories [Musa and Eriksson, 2012], the relationship between them [Cheng et al., 2012, Barbera et al., 2013, Cunche et al., 2014] and predict which of them will be most likely to meet again [Cunche et al., 2012]. Using the semantic information present in these probe requests, such as names of previously connected APs, it is possible to understand the nature of these users at a large scale [Di Luzio et al., 2016]. Using the received signal strengths from pre placed devices we can monitor the presence and movement of entities that are not even carrying a Wi-Fi enabled device [28].

Because of the security and privacy risks posed by the Wi-Fi protocol's use of hardware based MAC address, various methods to strengthen the security have been proposed [Pang et al., 2007b, Greenstein et al., 2008]. The randomisation of MAC addresses has become more mainstream in mobile devices with the introduction of it as a default operating system behaviour in iOS 8 by Apple Inc. Since MAC randomisation is not a perfect solution [Mathieu Cunche, 2016] there have been numerous attempts to fingerprint unique devices from the randomised anonymous information present in the probe request frames for the purposes of

[26] Julien Freudiger. How talkative is your mobile device?: An experimental study of Wi-Fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '15, pages 8:1–8:6, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3623-9. DOI: 10.1145/2766498.2766517. URL http://doi.acm.org/10.1145/2766498.2766517

[27] Mathieu Cunche. I know your mac address: targeted tracking of individual using wi-fi, 2014. ISSN 22638733

[28] A Elgohary. On detecting device-free entities using wifi signals. *ece.uwaterloo.ca*, 2013. URL https://ece.uwaterloo.ca/{~}aelgohar/stat841-report.pdf{%}5Cnpapers3://publication/uuid/D6821814-0041-47E6-9A26-96A32F41B07F

trajectory tracking and access point security. The methods used are decomposition of OUIs where detailed device model information is estimated by analysing an already known dataset of OUIs [29]; Scrambler attack where a small part of the physical layer specification for Wi-Fi is used [Bloessl et al., 2015]; and finally, the timing attack where the packet sequence information present in the probe request frame is used [Matte et al., 2016, Cheng and Wang, 2016]. A combination of these methodologies has been proven to de-anonymise randomised MAC addresses [Vanhoef et al., 2016]. In addition to tracking, Wi-Fi probe requests can be aggregated to uncover the urban wireless landscape [Rose and Welsh, 2010] and used to reveal human activity at large scales [Qin et al., 2013], pedestrian numbers in crowds [Schauer et al., 2014, Fukuzaki et al., 2015] and also counting people in hyper local scales such as queues [Wang et al., 2013]. With enough infrastructure we can aim to generate a real-time census of the city [Kontokosta and Johnson, 2016] and also predict the amount of time a device will spend around the sensor as well [Manweiler et al., 2013]. Similar to GPS data this can be used as an additional control layer for interpolation techniques such as map merging [Erinc et al., 2013].

[29] Jeremy Martin, Erik Rye, and Robert Beverly. Decomposition of mac address structure for granular device inference. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pages 78–88. ACM, 2016

2.3.7 *Consumer data*

In addition to the direct data from the sensors themselves the content generated from the mobile devices such as social media data or smart-cards [30] can provide a viable proxy for estimating the level and nature of human activity. The use of geo-located tweets on the study of small-area dynamic population estimation [Ordonez and Erath, 2012, Marchetti et al., 2015, McKenzie et al., 2015, Lansley and Longley, 2016b], geo-demographics [Bawa-Cavia, 2011, Longley et al., 2015, Lansley and Longley, 2016a] and global mobility [Hawelka et al., 2014] has been thoroughly explored. These data sources are shown to be useful in social sciences [Crane and Sornette, 2008], abnormal event detection [Chae et al., 2012] and analysing urban environments [Sagl et al., 2012]. It can also be used as a control layer for interpolation techniques we discussed earlier [Lin and Cromley, 2015].

[30] Chen Zhong, Michael Batty, Ed Manley, Jiaqiu Wang, Zijia Wang, Feng Chen, and Gerhard Schmitt. Variability in regularity: Mining temporal mobility patterns in london, singapore and beijing using smart-card data. *PLoS ONE*, 11(2), 2016. ISSN 19326203. DOI: 10.1371/journal.pone.0149222

2.4 *Research Gaps and Opportunities*

In this section we summarise the previous sections to find out the best possible technology for further research and discuss the research gaps and opportunities available to us. Table 2.1 summarises the above discussion to evaluate all the relevant technologies that can be used for the data collection and analysis for the study of human activity at a granular level.

Technology	Interpolation	Bespoke	Cellular	GPS	Wi-Fi
Coverage*	Local	City	All	Local	All
Certainty*	Very Low	High	Medium	High	Medium
Independence*	Low	Very High	Low	Medium	High
Intrusiveness*	Low	Medium	High	High	Medium
Granularity*	Very Low	Very High	Medium	High	High
Ease of Collection*	Medium	Low	Medium	Low	High
Scalability*	Medium	Low	High	Medium	High

Table 2.1: Evaluation of different technologies or approaches that can be used for data collection.

** coverage - the density and extent of the current infrastructure. Certainty - the lack of uncertainty in the data. Independence -How much the technique depends on secondary data. Intrusiveness - the potential for infringement of users' privacy. Granularity - the smallest spatial and temporal at which data could be collected. Ease of Collection - how efficient it is collect data in terms of time and resources. Scalability - the potential for the technology to improve coverage.*

We can observer that Wi-Fi offers the best possible technology in terms of flexibility and scalability for data collection through mobile devices at an individual level while posing some risk to privacy of participants and involves uncertainty regarding the field of measurement.

[Pinelli et al., 2015] looks at a comparison of various approaches of collecting and analysing mobile phone location data. The research identifies two major approaches in collecting device location data - Event-driven and Network-driven. The event-driven approach is centered around the mobile devices generating data through their day to day activities. The major sources of event-driven data are Call Detail Records(CDR) and internet use. Network-driven approach is centered around the service provision infrastructure such as cellphone towers, Wi-Fi base stations etc. The methods used to collect network-driven data are periodic update - where the device sends an update stating the base station it is connected to, handover - where the device information is recorded as they are moving between base stations and location update - where the location of the device is recorded based on the base stations it is connected to. The research used a set of anonymised mobile phone location data from a Belgian telecom operator for the city of Mons from which various event-driven and network driven scenarios were simulated. The authors compared these simulated scenarios for application-independent and

application-dependent cases such as spatial dispersal, classes of users, count estimation and flow estimation to understand their relative advantages and disadvantages. Through these comparisons it was shown that using network-driven mobile phone location data is more advantageous compared to the widely used event-driven ones.

From the literature search we can summarise that there is a considerable opportunity in the collection and analysis of mobile phone based data for measuring hyper-local, spatio-temporal dynamics of human activity. The potential for research gaps are discussed in detail in the following sections.

2.4.1 *Ambient population analysis*

Opportunity 1: *Design and collection of national/regional, longitudinal, grass root level data set which enables study of population both spatially and temporally.*

Previous research in this area of study has been limited to either national/ regional level studies using centrally collected residential population data such as censuses or to area level studies conducted with mobile devices based technologies. Though there were some efforts in collecting and using mobile phone data at national/ regional level we have never been presented with such unprecedented level of data available now.

For example, [Qin et al., 2013] demonstrate that it is possible to detect and quantify human presence at locations using probe requests with a detection rate of 86%. Along with the evaluation of the various algorithms for channel switching the research also successfully classifies these detected human presence into distinct activities in a non-intrusive way. Though this work predates both the MAC address randomisation and wide spread use of mobile experienced these days, the explosion of consumer data both publicly available and privately held presents previously unseen opportunity and also limited by the privacy concerns that arise with them. There is an immense opportunity to collect and standardise a large national level dataset which closely linked to the population distribution and movement in an anonymised way which then can be used to understand the distribution of population and its change. There is a need to extend such effort longitudinally which can give us insights in to the change of such phenomenon in time. This has the potential to enable us to ask broad questions such as,

- What are the trends in the footfall in UK?
- What are the daily rhythm of different cites?
- How much a weather event affect economy of a region?

Such dataset, in conjunction with other consumer data sources, in

addition to augmenting each other to improve their quality, can vastly improve our understanding of the structure and dynamics of population.

2.4.2 *Device fingerprinting*

Opportunity 2: *Developing models and methods to identify anomalies in the data and underlying events causing them*

The privacy concerns about the data collected from personal mobile devices has pushed the industry and users to find ways to anonymise the data generate over the last decade. All the mobility studies recording user trajectories across space and time are rendered infeasible with the cryptographic hashing and randomisation techniques employed by the devices. This along with progressive legislation such as General Data Protection Regulation have severely constrained the data available for mobility research. As we see later, even the estimation of ambient population is limited by these developments.

[Vanhoef et al., 2016] presents several novel methods of abusing the features of the Wi-Fi standards to track mobile devices even when the MAC addresses were randomised. This research shows the possibility of using the information elements present in the probe requests along with the sequence numbers to fingerprint the mobile device which sent the request with an accuracy of the 50% within a 20 minute interval with a possibility of improvement with known scrambler 'seeds' - the randomisation factor used by popular commercial devices. Though this sounds promising for short intervals, since this research, manufacturers have stopped including non-mandatory information elements which can affect the accuracy significantly. The research also features two other methods to reverse engineer the original MAC addresses from the randomised ones - first where known hotspots were spoofed to trick the mobile devices in revealing their real addresses and the second where a different protocol requests were used. Both these methods cannot be used extensively since the former is not ethically sound and the latter is not widely used by all mobile devices.

Since the above study and the following ones were conducted from security perspective - evaluating the robustness of the randomisation/ obfuscation procedure, they focus on de-anonymising the obfuscated data to recreate the personal information from them while demonstrating vulnerabilities in the standard and associated risks for the users. In this context, there is a clear gap for research in to methods to rather carry out fingerprinting of these devices using patterns in the data to create useful information from them without actually de-anonymise the data. This can lead to production of data-sets and methodologies which will

enable us to,

- Get accurate estimation of ambient populations.
- Understand the movement of the population in space and time.

2.4.3 *Event Detection*

Having granular spatio-temporal data on population at an area level also enables us to look at the activity of people at this scale. For example, the spike in Wi-Fi activity at a certain area at a certain time can illuminate us with a specific event that is happening in that area. Thought research have been done on this area using social media data, a longitudinal data-set collected using mobile technology can enable us to formalise the models needed to identify anomalies, quantify the causation of such anomalies to real world event. [Kontokosta and Johnson, 2016] discuss the use of Wi-Fi data for a 'real-time' census of the city with a case study of New York City's Lower Manhattan neighborhood. The research collects around 20 million Wi-Fi data points during 2015 and presents a model to create real-time, on-the-fly population estimates with fine granularity. The research demonstrates the feasibility of the pursuit along with the potential significance of such localised population estimates for use within the domains of city operations and policy, strategic long-term planning processes, emergency response etc. There are opportunities to ask questions such as,

Opportunity 3: *Developing models and methods to identify anomalies in the data and underlying events causing them*

- How did the tube strike affected London?
- What were the hot spots for New years celebration?
- What was effect of a road closure in specific part of the city?

2.4.4 *Pedestrian Flow*

Similar to the device fingerprinting, estimating and understanding pedestrian flow in the street network has immense opportunities since the anonymisation of mobile devices has taken off. Even when the problem of the identifying unique fingerprints of users in the data has not been solved, there is a need to understand the overall performance of the street network in terms of pedestrian flow just from the precise, granular data available, especially when the data source is as unstructured and noisy as the Wi-Fi sensors.

[Musa and Eriksson, 2011] use the Wi-Fi probe requests collected in a 12-hour trial on a busy road to describe a passive tracking system for

mobile devices. The research proposes a trajectory estimation method based on Viterbi's algorithm which estimates the most-likely spatial path taken from the information on when and where they have been detected. Although the research extends this trial into a 9-month deployment and demonstrates trajectory estimates with high accuracy, the problem still remains where we need to extract trajectories of users without actually being able to identify them.

This problem can he approached in two ways,

Opportunity 4: *Estimating flow of pedestrians in the street network from Wi-Fi data*

1. Probabilistic approach - Where the relationship between the temporal change in volumes at locations are modelled. For e.g. how much and how often the footfall counts at one location mirrors/ follows other location gives us an idea of how many pedestrians move from one location to the other.
2. Interpolation - Where the relationship between the locations are defined in terms of multiple variables such as how similar they are, how close they are etc. These relationships can in turn used to build a graph of locations and use this graph as a source to interpolate other locations.

2.4.5 Nature and Change of Places

Though there are extensive research in using ambient population and people's movement to understand the form and function of the space, the mobile technologies have introduced the opportunity to remove the subjectivity from them. With access to highly granular and long-term data sets, we can hope to look into the how the places have changed over time and how the external factors such as policy and economy has affected them. There are opportunities to ask questions such as,

Opportunity 5: *Using long term data to detect the nature and change of form and function of a place.*

- How does UK's exit from EU affect its high streets?
- Has a specific area has become more or less vibrant?

3
Collecting Wi-Fi Data

From the literature review in Chapter 2, we observed that of all the technologies discussed, Wi-Fi seems to be the most promising one for our purposes. We observed the advantages of Wi-Fi based data collection as,

- Universality as a standard technology globally,
- Independence from other types of data sources or infrastructure,
- High level of granularity both spatially and temporally,
- Possibility of passive data collection,
- Extreme ease of collection in terms of cost and effort and
- Scalability to cover study large areas.

Though it has its pitfalls in terms of intrusiveness resulting in risk to the privacy of the users, as well as bias and uncertainty, Wi-Fi provides us with a strong base framework for fulfilling the opportunity to design and collect a large, long-term and granular dataset which can be used for studying human activity.

In this chapter, we continue our research by looking at Wi-Fi technology closely to understand how it can be used to achieve the aforementioned goal. We start by looking at the Wi-Fi specification [1] and focus on the information available within the Wi-Fi probe requests. We then design and implement a series of data collection exercises which collect probe requests in various location with increasing level of complexity for analysis. We explore these datasets briefly to understand the usefulness of each set of information present in the probe requests along with the uncertainties in them. We also introduce the 'Smart Street Sensor' project - a national scale effort for collecting Wi-Fi data at high streets across the United Kingdom. Finally we summarise the data collection procedure with a detailed look at the uncertainties in these datasets and draw conclusions for further lines of research into alleviating the uncertainty and

[1] IEEE. IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, Dec 2016

noise so that the datasets can be used to estimate human activity with confidence.

3.1 *Wi-Fi as a Source of Data*

Since the formation of 'Wi-Fi alliance' in 1999 to hold the trademark, Wi-Fi (Wireless Fidelity) has become synonymous with the IEEE 802.11 standard based internet connectivity. Today almost almost all devices use this standard to create and connect to local area networks wirelessly. Due to its high fidelity and immense throughput up to 1 Gigabits per second, Wi-Fi has become the choice of technology for wirelessly transferring large amount data through networks. The adoption of 'smart' mobile devices Smartphones across the world has further cemented Wi-Fi's position as one of the most ubiquitous technologies which many people use every day. In developed economies such as the UK, this has never been more true and having an infrastructure to serve and receive Wi-Fi signals greatly affects the ability to connect to the internet in many areas. With close to 87%[2] of the population carrying one or more of these smart devices with Wi-Fi capability, provision of Wi-Fi as a service has become essential for any place, thus making Wi-Fi (alongside mobile networks) one of the most used technologies to access the internet.

[2] Deloitte. Mobile consumer survey - united kingdom, 2018. URL https://www.deloitte.co.uk/mobileuk/

Though the end goal of internet connectivity is the same, Wi-Fi greatly differs from internet connectivity through mobile networks such 3G/4G. The first difference is the range of the network: unlike mobile infrastructure where a single tower can serve mobile phones for miles, Wi-Fi is designed to be an extension of the wired networking, thus creating short range network with a range of 20 meters. Due to this low-range and high throughput property, Wi-Fi is used primarily as a distributed infrastructure operated by owners of premises as a means to provide high speed connectivity to the users of these buildings as opposed to the large, national level, monolithic infrastructure that runs the mobile network. This creates a situation where urban areas are populated by hundreds and thousands of these small area networks to which any mobile device can connect to. Unlike the mobile service providers and their customers, these Wi-Fi networks and mobile devices don't trust each other with specialised hardware. This creates a need for an introduction procedure - a sort of handshake between them - whereby they exchange information about themselves. Moreover since these mobile devices constantly move across these Wi-Fi networks, it becomes necessary for them to carry out

these 'handshake' processes regularly and frequently so that they can traverse between the networks without loss of connectivity. This need for constant lookouts for new networks is solved by the 'Probe requests'.

3.1.1 *Probe requests*

There have been numerous iterations and versions of the IEEE 802.11 standards but essentially all of them operate by exchanging packets of information called 'datagrams' or 'frames'. These frames have the information that is being exchanged along with the meta data and information on the device that is sending them. Some of these frames have special purposes: one such purpose is the 'network discovery'. The frames used for this purpose by the mobile device and the Access Point (AP) are called the 'probe request' and 'probe response' respectively. Though the actual information exchanged between these devices are usually encrypted, these probe requests are unencrypted and are accessible to any device which is listening. The structure of a probe request is shown in Figure 3.1.

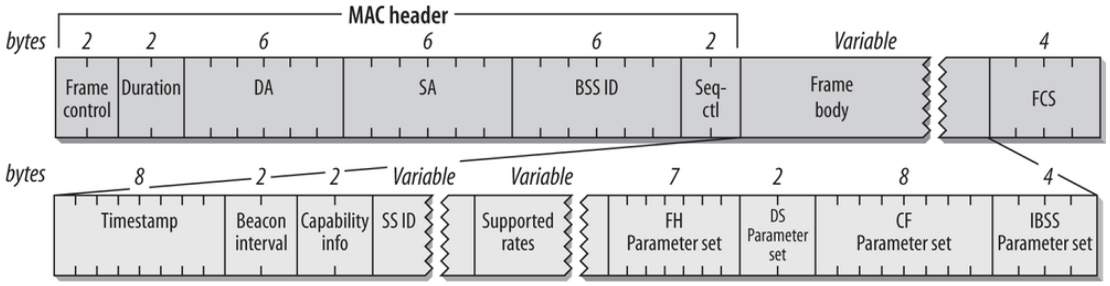


Figure 3.1: Structure of a probe request frame.

Source: IEEE 802.11 specification.
DA - Destination Address. SA - Sender Address. BSSID - Broadcast or multicast address. FH - Frequency hopping. OS - Optional. CF - Contention free. DS - Direct Sequnce.

We can observe that the fixed part of the mandatory frames are in the front; these show the identity of the mobile device generating the frame along with the identity of the AP that is receiving the frame. There are additional meta data such as the sequence number of the frame, and controls denoting where the frame starts and ends. There are also a number of variable information which can be used to transfer data. For probe requests, the destination device is set as 'broadcast' and the variable part usually contains the payload. For probe request frames, this payload consists of 'information elements' which has data regarding the capabilities of the device organised in units known as 'tags' or 'parameter sets'. The significant information present in a probe request is detailed in Table 3.1 and a full list of information available from a probe request is shown in the form of a sample probe request in appendix 7.5

Essentially the above information is sent over and over by the mobile device which expects a reply from nearby APs so that it can keep a list of networks it can connect to. This process is usually carried out even when the Wi-Fi is switched off in the operating system so that the connection times are faster once it is switched on. Moreover operating systems use the replies they get for these probe requests and triangulate the device location with respect to the APs with location information on AP's collected through surveys or crowdsourcing, thus acting as a quick and easy localisation solution which along with the above makes this probing process almost non-stop.

Table 3.1: Significant information included in a probe request

Field	Notes
Source Address	Media Access Control (MAC) address
Time stamp	Precise time at which the frame is received
Received Signal Strength (dBm)	The strength of the received signal
Length of the frame	Total length of the frame in bytes
Duration of transmission	Time it took to transmit the frame in milliseconds
Information Elements	List of various information about the device
Known Networks	Name of networks that are already known to the device

3.1.2 MAC address

Media Access Control (MAC) address is a 6 byte unique identifier assigned to a device on a network. It is similar to the Internet Protocol (IP) address but assigned at the interface controller level by the manufacturer of the device. Although the IP address of a mobile device might change regularly, the MAC address usually remains the same for the lifetime of the device making it akin to a unique identifier of a device and therefore highly significant. The MAC address has two parts: the first 2 bytes are known as the Organisationally Unique Identifier (OUI) and gives us information about the manufacturer of the network card. Organisations need to register with IEEE to be assigned an OUI which they can use to generate a full MAC address; the second 2 bytes are unique to device itself. Together both form the full MAC address which is unique to every device globally. The biggest draw for using Wi-Fi for mobility analysis comes from the fact that this globally unique identification is sent out regularly by mobile devices and can be collected passively through probe requests.

As we saw in our literature review, this also creates an immense risk in terms of infringement of privacy both for the manufacturer and the user. Manufacturers of critical hardware components who do not want

their unique MAC address to be publicised usually opt for registering a 'private' OUI which will be neither given out to other manufacturers nor published publicly. Users (their mobile devices) who don't want to be tracked using their MAC addresses use a temporary MAC address which is unique only to the local network - a 'local' OUI rather than using a 'global' OUI for unencrypted communications and switch to their original MAC address when a trusted encrypted connect has been established. This lack of uniqueness can be inferred from the second character of the MAC address being E, A, 2 or 6. Though this provides reasonably better privacy to the mobile users it also limits our ability to use the MAC address from the probe requests as in previous studies conducted with Wi-Fi. It is important to note that this is not a security measure, but rather an exception made available by IEEE 802.11 for assigning temporary addresses in ad-hoc networks which has been used by most modern operating systems.

Essentially, there are two types of MAC addresses based on whether they have a public OUI or a private OUI. This distinction does not affect their uniqueness or usefulness in mobility research but hinders us from knowing about the device from the MAC address. There are also two types of MAC addresses based on them being unique globally or just in the local network. This distinction affects the feasibility of using the MAC addresses for device tracking or for studying movement of the users.

To summarise the above, we looked at the IEEE 802.11 standard to examine the significance and nature of the 'probe requests' which are constantly broadcast by mobile devices. We identified information present in these probe requests which is relevant to our study and examined the uniquely identifiable MAC address field in detail. We found that though a MAC address provides a way to globally identify a mobile device from the probe requests it generates, this field can often be masked by using locally assigned addresses. We also observe that there is other relevant information which, when combined, can provide us with an alternative to solely using MAC addresses.

3.2 *Initial Experiments*

With our theoretical understanding of the Wi-Fi standard and its capabilities, we move on to looking at the Wi-Fi landscape in the real-world. We achieve this by designing small independent experiments where we

record the Wi-Fi probe requests within controlled conditions along with the knowledge of the ambient population of the field of measurement. We then look at the collected probe requests, examine them in detail to look at their properties, aggregate them to footfall counts and compare them with the real-world counts to get an overall idea of how well they translate into real counts. The aim of these experiments is to know more about the probe requests data and pick out the uncertainties and opportunities present in them. The objectives here are,

1. Design a simple method to collect probe requests.
2. Select locations with different levels of complexity.
3. Collect real-world data through manual counting.
4. Analyse the probe requests to extract useful information.

3.2.1 *Experiment Design*

The first step was to design a simple method to collect Wi-Fi probe requests. We accomplished this by using the open source, free software *tshark* [3] on a regular laptop. First, we put the Wi-Fi module of the laptop in 'Monitor mode' where it behaves as a wireless access point rather than a receiver. Then we invoke the command line interface of the Wireshark programme tshark to collect the Wi-Fi probe requests received by the laptop in Character Separated File (CSV) format in the file system. The full shell script which collects the data is given below,

[3] Gerald Combs and Contributors. Wireshark - network protocol analyzer. https://www.wireshark.org/about.html, 2018

```
#! /bin/bash
tshark \
  -I -i en0 \
  -T fields \
  -E separator=, \
  -E quote=d \
    -e frame.time \
    -e frame.len \
    -e wlan_radio.signal_dbm \
    -e wlan_radio.duration \
    -e wlan.sa_resolved \
    -e wlan.seq \
    -e wlan.tag.length \
    -e wlan.ssid \
  type mgt subtype probe-req and broadcast
```

It is important to note that this script only collects the particular data from the probe requests which we found to be relevant to our needs. The fields marked with *-e* are the ones which were collected and they correspond to the information in the probe requests as follows,

1. *frame.time* - Time stamp when the packet was received in microseconds.
2. *frame.len* - Total length of the packet in bytes.
3. *wlan_radio.signal_dbm* - Strength of the signal which delivered the probe request in dBm.
4. *wlan_radio.duration* - The duration for which the signal has been transmitted.
5. *wlan.sa_resolved* - The MAC address of the source device where the first part is resolved into a vendor name concatenated with 6 characters of the device part.
6. *wlan.seq* - Sequence number of the packet assigned by the source device.
7. *wlan.tag.length* - A list of lengths of the tags attached to the packet this acts a signature of the information contained within those tags and
8. *wlan.ssid* - The network for which the probe request is being sent for. This information is optional.

The name of the manufacturer/ vendor of the Wi-Fi module is extracted from the *wlan.sa_resolved* field into a separate column and the original field is hashed using the SHA256 algorithm [4] implemented in R. In addition to this, the pedestrians next to the sensor were counted manually by the surveyor.

[4] Shay Gueron, Simon Johnson, and Jesse Walker. Sha-512/256. In *Proceedings of the 2011 Eighth International Conference on Information Technology: New Generations*, ITNG '11, pages 354–358, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-0-7695-4367-3. DOI: 10.1109/ITNG.2011.69. URL http://dx.doi.org/10.1109/ITNG.2011.69

3.2.2 *Living room*

The first set of experiment was conducted with the laptop in the researcher's living room. The primary aim of this experiment to collect an initial set of probe requests is to understand the information present in them in detail. The living room had 2 Wi-Fi enabled devices - an Android phone manufactured by Motorola and an Android TV box manufactured by Remix. The other rooms in the house had an iPhone from Apple running iOS9 and an Android phone from Samsung in the rooms next door. The script was left running on the laptop on 15 Nov 2015 from 19:44 to 21:15 with an unexpected failure of approximately 15 minutes in between from 19:55 to 20:10 approximately. In this duration, we collected around 3000 probe requests at the rate of 38 requests per minute.

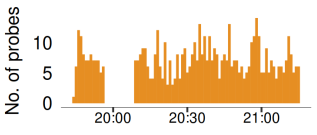


Figure 3.2: Number of probe requests collected every minute on 15 October 2017

The first thing we tried with the probe requests was to aggregate them based on their MAC addresses. Before mobile devices started randomising their MAC addresses, this should have accurately reflected the number of devices around the laptop. The data when aggregated showed that there were around 211 unique MAC addresses recorded. Being a residential area far away from traffic, these MACs are most likely not from unique devices. The high number must be the result of randomisation. Moreover, since we know that there are only 2 - 4 devices in the house, there must be noise from significant distances beyond the house. The number of unique MACs recorded every minute are shown in Figure 3.2. We observe that on average we captured around 7 unique MAC addresses every minute which is quite far from the 2-4 range we were looking for.

Having established that just the MAC addresses are not enough to accurately translate the probe requests into the number of devices around the sensor, we started to look at the other information we collected from the probe requests alongside the MAC addresses. First, we tried to isolate all the randomised, local MAC addresses by looking at the resolved vendor part. We aggregated the probe requests based on the vendor name present in the MAC addresses against all the other information present in them. The results are shown in Table 3.2 We looked at how many unique values were present in these fields compared to the total number of probe requests. Initially we assumed that the randomised probe requests won't have public OUIs which and hence the probe requests which can be resolved should be the real addresses. But when we looked at the probes to MAC ratio of Google and Compex we realised that even the local MAC addresses could be registered public. This showed even when the OUI has been resolved into a vendor name, the original needs to be preserved for analysis. Samsung is a special case since we know from the specification that whilst their devices do not randomise the addresses, they also have many unique addresses which need to be taken into account.

We observe 24 different vendors in the data. Even if we assume one device per vendor, it is impossible for the sensor to pick up 24 different devices without a significantly larger area of measurement than we expected. We need a way to filter out this noise which is generated from the edge of the field of measurement. This is where the signal strength shows good promise. Looking at the Table 3.2 we can see that two of these vendors show significantly high average signal strength - Google

and Fn-LinkT, which can easily correspond to the two devices present in the room. This can be explained by the decay of the signal as it passes through the walls. In our simple example, we can filter out almost all the noise just by using the signal strength of the probe requests.

Table 3.2: Number of unique values present in each field captured from the probe requests aggregated by the vendor names

Vendor	No. of probes	MAC addr.	Signal (avg.)	Frame length	Dura-tion	Tags	SSID	Seq. no.
AmazonTe	101	1	-80.53	4	4	5	3	101
Apple	77	7	-86.29	4	4	9	4	77
ArrisGro	7	1	-91.71	1	1	1	1	7
Azurewav	215	4	-87.82	3	3	12	10	213
CompexPt	75	28	-88.17	3	3	5	29	74
CompexUs	4	1	-87.25	3	3	3	4	4
Dedicate	2	1	-92.50	1	1	1	1	2
Fn-LinkT	64	1	-60.58	2	2	6	1	64
Google	1347	76	-69.14	4	5	41	6	1157
HuaweiTe	11	3	-87.91	3	3	3	1	11
IntelCor	25	2	-84.04	3	3	4	3	25
LenovoMo	1	1	-93.00	1	1	1	1	1
LgElectr	1	1	-90.00	1	1	1	1	1
Microsof	3	1	-90.00	1	1	1	2	3
Nvidia	65	1	-82.91	2	2	4	2	65
OneplusT	3	1	-86.67	2	2	2	2	3
Pepwave	4	4	-90.00	1	1	1	1	4
Sagemcom	3	1	-88.67	1	1	1	1	3
SamsungE	655	27	-83.81	26	26	54	23	621
SonyMobi	56	2	-78.66	2	2	2	1	56
TctMobil	1	1	-90.00	1	1	1	1	1
Tp-LinkT	31	1	-86.16	1	1	3	1	31
Wisol	143	3	-71.91	4	5	6	3	142
XiaomiCo	3	2	-88.67	2	2	3	2	3
Unknown	110	40	-88.86	19	18	21	5	90

We then look at all the other information we collected from the probe requests and see how they compare to the MAC address for aggregating. We observe that frame length and duration provides better aggregation into unique values than MAC address when they are randomised, as seen with Compex and Google. Since the devices were essentially sending the same information repeatedly with just changed fixed-length MAC addresses, we expect that the same devices should be sending packets of the same length. We also observe that the duration of the transfer, being a function of the length of the signal, has almost the same amount of information in it as the frame length. We can confidently pick one of these fields and discard the other for further analysis. Though the tag lengths and SSID looked to be a promising way to uniquely fingerprint the devices they don't have enough volume in them to offer substantial advantage. The set of tag lengths are not as unique as the lengths or duration while the SSID information is sparse for most of the vendors. For example, 66% of probes with local OUIs, 50% of the ones with Google and 38% with Samsung don't have any SSID information in them. This makes them very poor candidates for useful information in aiding us in

finger printing unique devices.

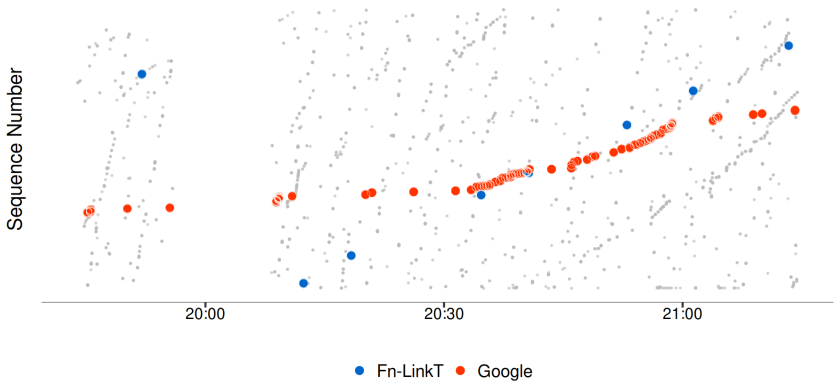


Figure 3.3: Sequence numbers plotted against timestamps showing clear patterns corresponding to unique devices.

Grey dots are probe requests with signal strengths lower than -70dBm.

Finally, we found that the sequence numbers are the most interesting part of the data collected. Although they don't uniquely identify the devices directly through aggregation, along with timestamps they do form visually discernible, interesting patterns that correspond to the mobile devices that generated them. In Figure 3.3 we have isolated the two vendors identified earlier (Fn_LinkT and Google) and filtered only the probes requests with signal strength of more than -70dBm. We then plotted their sequence numbers against the precise time stamps when they were received. We can clearly see two devices which were present in the room, which demonstrates the usefulness of the sequence number in estimating the actual devices around the sensors. We need to devise a method for isolating the 'tracks' left by the devices in terms of their sequence numbers over time. We can also observe the rotation of sequence numbers at 4096 for the Fn_LinkT device which needs to be considered while devising such method. Figure 3.4 shows a similar exploration of Samsung devices. Though from the table 3.1 it looks as if Samsung devices are randomising their MAC addresses, we can clearly see in the figure that there are only two devices which were present for a long time around the sensor and neither randomised their addresses. The diversity of MAC addresses were indeed unique devices which must have been located far away from the sensor.

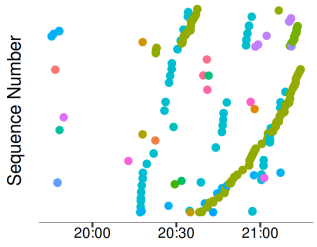


Figure 3.4: Sequence number patters in Samsung devices showing the diversity of MAC addresses showing that they are not randomised.

The colours show unique MAC addresses.

To summarise, we found that even when an unique non randomised MAC address is present when collecting Wi-Fi probe requests, we get significant noise from outside the perceived field of measurement. We also found that signal strength is a really good clue to filter out this noise. The frame length and duration looks promising for the same purposes, but they ultimately have the same information and can be

used interchangeably with similar results. Finally, we found that tag lengths and SSID are not useful information since they are either too varied or too sparse. Although the results of this exploratory analysis have been positive, the main challenge is to make sure these methods are feasible when dealing with more real-time, real world data. We need to devise a more real world experiment to see frame lengths and signal strength work in a bigger dataset for filtering out the noise.

3.2.3 *UCL South Cloisters*

This experiment was conducted collect a broader dataset from a real world setting so that we can examine the results from the previous experiments with further confidence. The specific goals were to validate the findings on signal strengths in respect to the distance from the sensor in the previous data, and to further examine the usefulness of the frame length parameter. We also wanted this to be a standard dataset on which we can test our methodologies before they are applied to a broader project such as Smart Street Sensors. The data collection was conducted in one of the corridors in UCL - Southern Cloisters - which attracts a lot of pedestrian traffic during term time. This corridor also has substantial seating areas along the side where students often sit down for long periods of time to work. This provides us with a source of devices which dwell near the sensors as they constantly sending out probe requests. This area is also used heavily for lunch and for exhibitions/ events attracting a large amount of visitors, thus making it ideal for 'stress testing' our methods for cleaning and aggregation. The position of the sensor with respect to building is shown in 3.5. The data were collected from 15:37 to 16:20 on 04 December 2017, a period during which we collected around 14,750 probe requests using the scripts mentioned earlier. We also manually counted 652 pedestrians walking directly in front of the sensors.

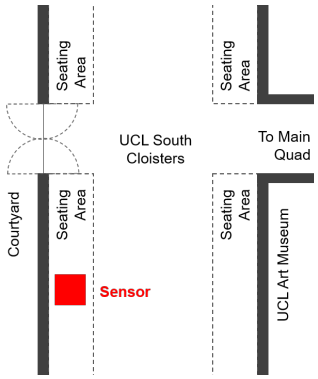


Figure 3.5: Illustration showing the configuration of the sensor at UCL south cloisters
* *Not to scale.*

Unlike the previously collected data in this experiment, we made sure that the OUI information is preserved even after resolving them to vendor names. With this information we looked at the second character of the OUI and categorised the probe requests as either 'local' - randomised, or 'global' - non randomised. We then compared them to the vendor names to find out if any manufacturers other than Google have registered OUIs in the local range. Figure 3.6 shows the distribution of vendors within both the local and global range of OUIs in terms of the number of probe requests collected.

We observed that 'Google' is the only registered public OUI found in the public range. We also noticed that the percentage of global MAC addresses collected was unusually large - 82%. This can be explained by the behaviour of the Apple devices while randomising the MAC addresses. Apple phones are known to randomise their addresses while probing for access points only when they are not connected to a Wi-Fi network already as documented by Vanhoef et al. [2016].

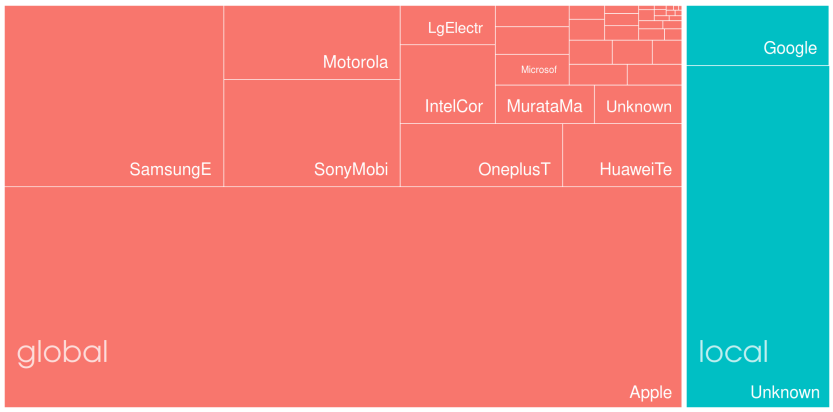


Figure 3.6: Composition of probe requests in terms of the vendor names and their type

Based on the number of probe requests

Since most of the members of UCL have access to the 'eduroam' network and are connected to it whilst on the campus, most of the Apple devices we captured haven't randomised their addresses. This made this dataset heavily biased and not suitable for testing device finger-printing methods, but it does give us an opportunity to examine the nature of probe requests generated by Apple devices in particular.

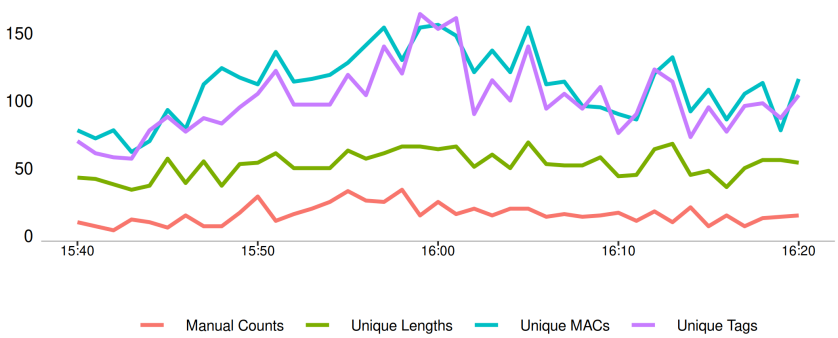


Figure 3.7: Comparison between the manual footfall count and aggregated counts from sensor collected data at UCL South Cloisters.

The second step was to see how much the sensor collected counts differ from the manual counts. We aggregated the sensor counts for every minute in terms of the number of probe requests, unique MAC addresses, and the unique frame lengths, and compared them to the manual counts done for each minute. The results are shown in Figure 3.7. We can

observe that the original Mean Average Percentage Error (MAPE) when aggregated with MAC addresses is around 736% showing the immense amount of noise we can experience in a real world environment. This was reduced to 643% and 300% when aggregated by tag lengths and frame lengths but it is still far from being anywhere near accurate for being able to be used for estimating footfall. When we filter the probe requests for just the ones which have signal strengths more than -70dBm - the threshold which we got from the previous experiment - the MAPE for aggregating by MAC addressed, tag lengths and frame lengths is reduced to 80%, 87% and 67% respectively. The results after filtering with signal strength are shown in Figure 3.8

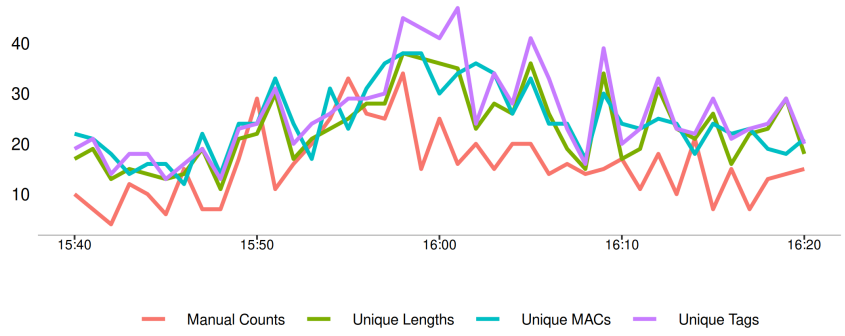


Figure 3.8: Comparison between the manual footfall count and aggregated counts from sensor collected data at UCL South Cloisters after filtering probes with low signal strength

Although the signal strength filtering works to remove noise, we are still not clear about how this works or what is the most optimum cut off for filtering. We looked at the distribution of the signal strengths to find that they do exhibit patterns in terms of concentration around certain cut-offs, as shown in Figure 3.9. These cut-offs can be detected dynamically from the data using one dimensional clustering methods such as k-means which are usually used to find the class intervals in one dimensional data. Figure 3.9 also shows the results of k-means clustering on the data to divide the data into 4 clusters.

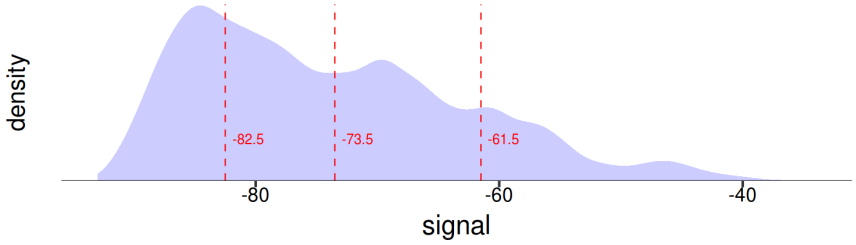


Figure 3.9: Density distribution of the signal strengths of the probe requests collected at UCL South Cloisters along with class intervals.

x-axis is measure in dBm as a proxy for distance. The class intervals calculated using k-means clustering with the number of clusters defined as 4.

To summarise, in this experiment conducted at UCL South Cloisters we collected a bigger set of data over a longer period of time to validate

the previous findings and to serve as test dataset for further research. We found that signal strength is one of the key pieces of information with which we can remove the external noise from the dataset. We also found that although the tag lengths and frame lengths look promising as a filter, they do not gives us any significant advantages. Unfortunately, the data were also found to have major bias towards non-randomised probe requests because of the availability of the campus Wi-Fi. This requires us to collect a more representative dataset for further research into using sequence numbers to finger print devices. Finally, we also found that accurate manual measurement of real footfall is challenging, and we need a better method to collect data for the surveyors in order to maintain accuracy and precision.

Figure 3.10: Location and configuration of Wi-Fi data collection carried out in Oxford Street, London.

3.2.4 *Oxford Street*

From the results of the previous two experiments we arrived at the task of devising a final 'real world' experiment collecting probe requests at a high street with high volume of footfall. Similar to the previous data, the aim here was to generate a dataset which can be used to test and validate signal strength based filtering and sequence number based clustering methodology against the scale and complexity of a busy, open public area such as a retail high street. The location chosen was Oxford street, London - one of the busiest retail streets in the world. The data was collected from 12:30 - 13:04 hrs on 20 December 2017 using the same methodology as above from a laptop in a backpack. The surveyor positioned himself at the front of a store while carrying the backpack and counted the people walking by the store on the pavement (3m wide approximately) using a mobile phone. The sensor was kept as close to the store window as possible, and the manual count was done as a cordon count in front of the store. The location where the data was collected and the configuration of the sensor with respect to the street is shown in Figure 3.10.

The manual counting was done using a node-js base command line app running under Termux on an Android phone. The application is detailed in section 7.1.1 which counts the number of times a key has been pressed on the phone. This has an additional advantage as the phone used is kept unconnected to any Wi-Fi and with the screen on for counting, emits probe requests at regular intervals. Moreover, we know the phone to be of the vendor 'Google' which randomises the MAC address, giving us a good base line to compare our results to.

The Wi-Fi sensor captured approximately 60,000 probe requests during the half hour period; 3,722 people were manually recorded walking on the pavement during that time. Initial exploration of the data is shown in Figure 3.11, where we compared the sensor aggregated counts to the manual counts of footfall. It shows that the data has a large amount of noise making it a suitable candidate for testing. Moreover, with 55% of local MAC addresses, it is free from a high concentration of global MAC addresses as we saw in the data from UCL corridors. This dataset is extensively used in the development of the filtering and cleaning methods and which are discussed in detail in Section 4.2.

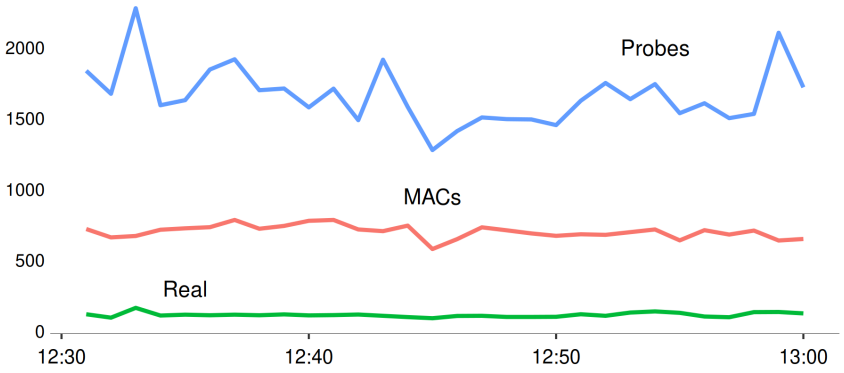


Figure 3.11: Comparison of the counts from aggregated probe requests and MAC addresses with manual counts at Oxford street, London.

In this section, we saw the design, implementation and initial results of small experiments we conducted to understand the nature of the probe requests and the opportunities they provide us with. We identified useful information in the probe requests and discarded the ones which were not useful The major conclusions arising from these experiments are,

1. The MAC address on its own is not enough to aggregate probe requests into devices or footfall.
2. The signal strength is crucial to removing noise from outside the field of measurement
3. The sequence number is promising in isolating devices when their MAC addresses are randomised.
4. Frame lengths, duration, tag lengths and SSID information do not add additional value in cleaning the data.

We finally collected a fairly representative Wi-Fi dataset from a high volume retail location for use in further research on methods to clean the data.

3.3 Pilot Study

As we see later in Section 4.2 the efficiency of the methods to clean and aggregate data not only depend on the noise and bias in the data itself but also on external factors such as, the configuration of the sensor in relation to the environment, the day of the week etc. Although the dataset captured in our initial experiments acts as a good starting point, it cannot enable us to generalise our findings to all possible configurations. This necessitates an even larger dataset to be collected over longer durations in the kinds of challenging situations that we would usually find in real world conditions. This was our primary motivation in conducting a pilot study collecting data at 5 locations across London. The aim was to collect probe requests with information we found relevant in the initial experiments for every location surveyed for at least a full week so that we can understand any patterns caused by the periodicity of the data. We also wanted to collect data at all of these locations in parallel for at least a week so that they can be compared to one another.

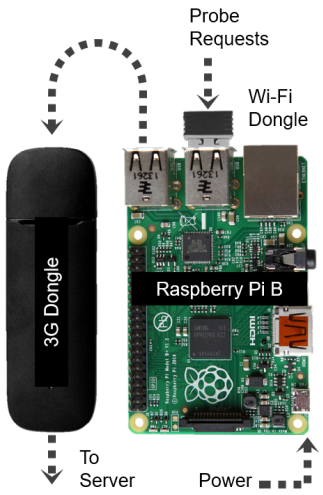


Figure 3.12: Hardware setup used to collect data in the pilot studies.

3.3.1 Methodology

The hardware setup for the sensors is illustrated in Figure 3.12. The design of the hardware is not original as it is heavily influenced by the proprietary technology of the data partner for the Smart Street Sensor project, albeit in a much simpler form. The core of the hardware is the general purpose single board computer Raspberry Pi Model B running Linux operating system. Two communication modules - 3G and Wi-Fi were connected to this machine via Universal Serial Bus interface. The 3G modem was equipped with a SIM card which it uses to connect to the internet, while the Wi-Fi modem is set to 'Monitor' mode. The board takes power from an outlet and the software is pre-installed with the operating system which resides in a Memory card.

The software used for the sensors consists of two parts - sensor software and server software. The sensor software was written as a mix of Bash script and NodeJS. Essentially these scripts use the Wireshark program to capture packets, parse them, anonymises the MAC address fields, add the location information, encodes them into JavaScript Object Notation format and finally sends it to a server through Web-Socket protocol. The code used at the sensor side is detailed in Appendix 7.2.1. At the server side we have a similar NodeJS application which listens to the data sent over web sockets, parse them and saves them to a PostgreSQL

database. The server side code is detailed in Appendix 7.2.2 and the schematic diagram for the whole process is shown in Figure 3.13. The information collected from each probe request at these locations are,

1. Time stamp at which it was received
2. MAC address of the source device.
3. Signal Strength of the packet.
4. Total length of the packet.
5. Sequence number of the packet.
6. OUI part of MAC address.
7. Location at which it is collected.

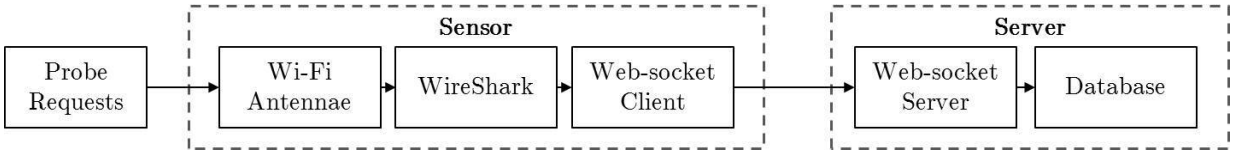


Figure 3.13: Schematic diagram showing the data collection process in the pilot study.

The manual counting at these locations were done using a custom application Soundararaj [2018]. The application was built for recording pedestrian footfall with precision and accuracy which was not possible when counted without the application. The app records the precise time stamp of every footfall with the precision of micro seconds which can be aggregated later at different time intervals. The code for the app is detailed in Section 7.1.2.

3.3.2 *Locations*

Five retail locations were chosen in consultation with the data partner for the pilot study, keeping in mind their complexity and volume of footfall. The sensors were installed at the locations in a phased manner and multiple manual counts were conducted at each location for 30 minute intervals. The locations and their descriptions are summarised in Table 3.3.

- *Location 1* is on the Camden high street in front of a mobile phone shop behind a bus stop. This location was chosen specifically because of the large amount of dwelling population at the bus stop and the stationary mobile devices inside the shop, both of which are expected to create a large amount of noise alongside the high footfall in the high street. The challenge here is to isolate the footfall from the two sources of noises which are at equal distance from the sensor.

- *Location 2* is on a square with a very low footfall but has a large amount of seating in the restaurants all around it. The challenge here is similar to that of the previous location in terms of noise, but just that the volume of actual footfall is low which makes it one of the hardest locations for accurately estimating footfall.
- *Location 3* is in front of Holborn station entrance in an information kiosk. This location was chosen for the really high volume of footfall from the station which is expected to cause noise. The challenge here is to be able to isolate the crowd inside the station from the footfall on the pavement.
- *Location 4* is in a fast-food restaurant in a shopping centre. The sensor has restaurant seating on one side and a pedestrian footfall on the other. The challenge here is that the stationary noise and the footfall are equidistant from the sensor.
- *Location 5* is at the frontage of a shop on the Strand with a mobile shop next door. This is the 'cleanest' location of all with only one clear source of noise which is at different distance from the footfall.

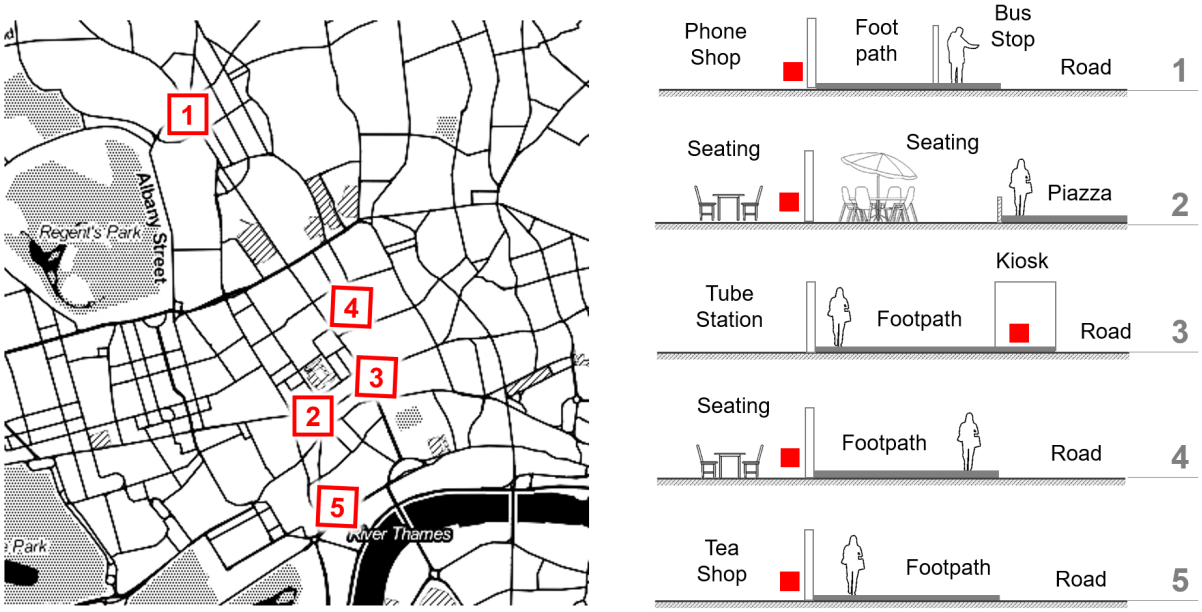


Figure 3.14: Pilot study locations in London along with their corresponding sensor installation configurations.

The sensors were operational through out February and March 2018, while the manual counts were conducted at these locations in half-hour sessions on at least two different days. The schedule of the data collection and the days at which the manual counting was done is shown in Figure 3.15. The survey was conducted for almost 2.5 months and about 33.5

million records were recorded which takes up to 1.8 GB of space on disk when encoded as text. During the manual counts around 10,000 people were counted walking past these sensors.

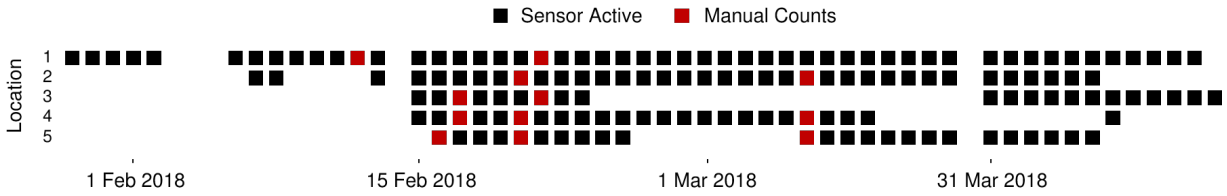


Figure 3.15: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data.

A detailed account of the volume and velocity of data collected at these locations was given in Table 3.3. The dataset collected was used extensively to develop and test the signal strength based filtering and sequence number based clustering methodology which are detailed in the Section 4.2.

Id	Location	Type	Installation notes	Probes*	Footfall**
1	Camden St.	Phone Shop	Bus stop in front	9.9 (297)	3683 (33)
2	St.Giles	Restaurant	Seating on both sides	3.9 (169)	0346 (05)
3	Holborn Stn.	Info. Kiosk	Front of station entrance	4.3 (303)	2956 (46)
4	Brunswick	Fast Food	Seating on one side	3.4 (210)	0960 (12)
5	The Strand	Tea Shop	Phone shop next door	8.4 (382)	1969 (21)

Table 3.3: Locations of data collection in the pilot study and the amount of data collected at each location.

* *Total probe requests in* $\times 10^6$*(per minute)* ** *Total footfall counted manually (per minute)*

3.4 *Smart Street Sensor Project*

The Smart Street Sensor project is one of the most comprehensive studies carried out on consumer volume and characteristics in retail areas across the UK. The project has been organised as a collaboration between the Consumer Data Research Centre, University College London (CDRC, UCL) and the Local Data Company (LDC). The project was designed to serve as the first and unique comprehensive research into the patterns of retail activity in high streets of United Kingdom by measuring their real-time footfall from collecting Wi-Fi probe requests. The data for the project was collected through sensors installed at around 1000 retail locations across UK.

The primary aim of the project is to improve our understanding of the dynamics of the high street retail climate in UK. As we saw in our literature review, unlike online retail, this involves the quantification and measurement of human activity at small scales, such as high streets, which is already the subject of active research. The key challenge in

this area is the collection of data at the smallest scales possible with minimal resources while not infringing on people's privacy. This challenge when solved can provide immense value to the occupiers of the retail premises who want to improve revenues, to landlords who want to increase the value of the property, to local authorities who want to improve the vibrancy of the retail economy, and finally to investors and consumers within the retail industry. The project also aims to facilitate decision making by these stakeholders, in addition to the tremendous opportunities for academic research.

Figure 3.16: Hardware setup used to collect data in the pilot studies.

3.4.1 *Methodology*

As a first step, various locations for the study were identified by the CDRC to include a wide geographical spread, different demographic characteristics, and range of retail centre profiles. Figure 3.17 shows all the locations in the United Kingdom city-wise and Table 3.4 shows the regional distribution of the locations.

Table 3.4: Regional distribution of Smart Street Sensor locations across UK

Region	Locations
Greater London	479
Scotland	118
Yorkshire and the Humber	114
South East	103
North West	98
South West	87
East Midlands	68
East Of England	49
West Midlands	39
North East	26
Wales	17
Northern Ireland	2

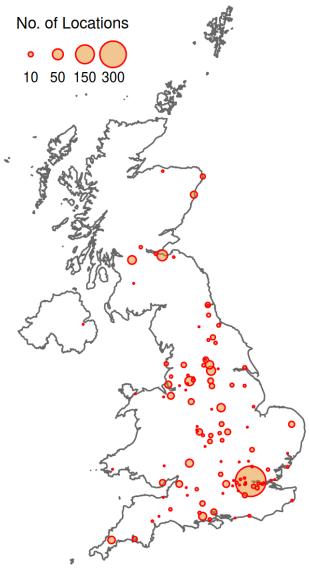


Figure 3.17: Distribution of locations with Smart Street Sensors installed.

We can see that the project has a strong London bias which along with other retail centres in the Greater London area, accounts for almost half of the locations. We must also note that the locations are retail and any insight from the data needs to be looked at with a retail point of view.

A custom footfall counting technology using Wi-Fi sensors (Figure 3.16) was also designed and developed by LDC, and the sensors were installed at the identified locations. The sensor employs proprietary hardware and software, and monitors and records the probe requests sent by Wi-Fi enabled mobile devices present in its range. In addition to this, the number of people walking by the sensor were counted manually for short time periods during the installation of the sensors at the corresponding locations. The project aimed to combine these two sets of data to use as a proxy for estimating footfall at these locations. The

potentially personally identifiable information collected on the mobile devices are converted into a unique cryptographic hash at the sensor level and the data is sent to central server via encrypted channel for storage. This data is then retrieved securely for the preparation of the commercial dashboards by LDC and for research purposes by CDRC.

The sensors are usually installed on a partnering retailer's shop window so that its range covers the pavement in front of the shops. A typical configuration of a sensor in a location with respect to the premises and the pavement in front of it is illustrated in Figure 3.18. There are also a small percentage (3%) of the devices which are installed within large shops to monitor internal footfall. Each device collects data independently and uploads the collected data to a central microsoft cloud facility (Azure) container at regular intervals of 5 minutes through a dedicated 3G mobile data connection. The sensor hardware has been improved over the course of the project, and currently has built in failure prevention mechanisms such as, backup battery for power failures, automatic reboot capabilities, and in-device memory for holding data when the internet is not available. The project began with the first sensor installation in July 2015, and has grown to an average of 650 daily active sensors as of January 2019, with a total of 1200 locations been involved in the project since its inception. We have collected around 2TB of data comprising of around 73 billion probe requests.

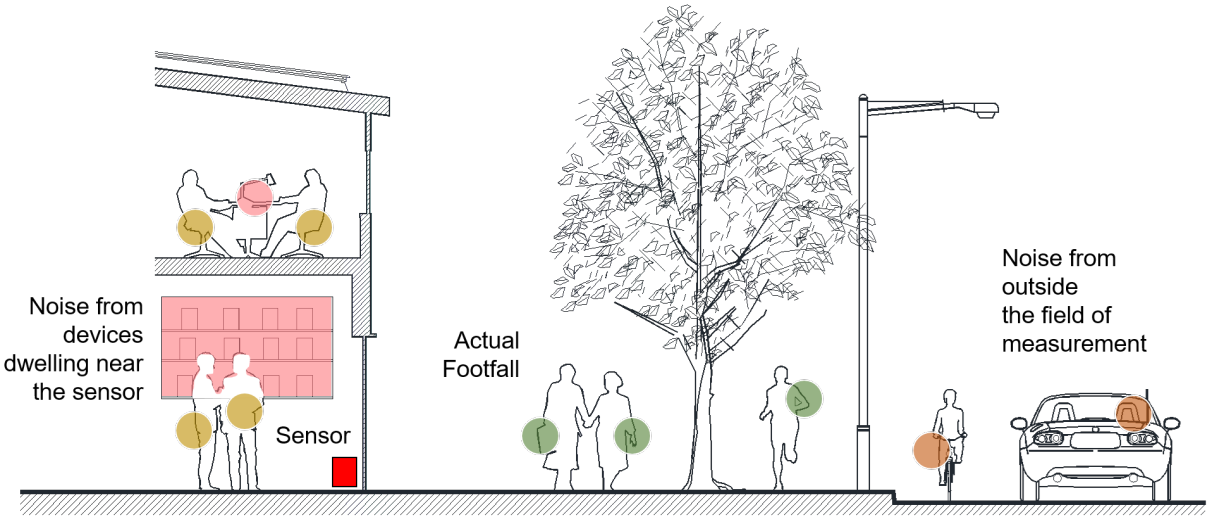


Figure 3.18: Cross section showing a typical installation of Smart Street Sensor in a retail frontage.

Due to the scale and the commercial nature of the project, the sensors collect fewer data per probe than the previous experiments. The information collected by the Smart Street Sensors are the 5 minute interval when the probe request was collected, hashed MAC addresses and signal

strength. The probe requests within the same five minute intervals are aggregated by the MAC address, hence the signal strengths are aggregated to the minimum signal strength reported. Due to the longitudinal nature of the project, the data collection methods have changed over time as well. The hardware was upgraded with more capabilities in early 2016, the interval they reboot at was adjusted several times in 2017, and finally, due to the MAC randomisation problem accentuated in the later part of 2017, the signal strength aggregation was changed from minimum to maximum in March 2018. Essentially, the data have changed over time and we need to consider the changes while devising the methodologies for cleaning the data.

3.5 *Discussion*

In the previous sections we designed and implemented data collection processes to arrive at 3 different datasets: Small Experiments, Pilot Study, and Smart Street Sensor project. The small experiments were designed as way to collect as much data as possible from the probe requests for short periods of time in order to collect small sets of comprehensive data under controlled conditions for exploratory purposes. The pilot study extended this further by collecting data for a longer time in real world conditions, aiming to validate the insights achieved with the experiments and the methodologies we devise for the research. The Smart Street Sensor project is the most comprehensive study which collects very small focussed set of data in a probe requestat a national level for very long periods of time. These datasets give us a well-rounded set of data to set up our toolkit and devise our methodologies. The summary of the datasets in terms of their characteristics is shown in Table 3.5.

Before we move on to develop methods to process the data into information on footfall in these locations, the crucial action is to look at the possible biases and uncertainties in these datasets arising due to the data collection methodology and from the broader context. These form the framework on which we built our data processing pipeline where we propose to solve each of these uncertainties in each step.

From our understanding of the data, we observe that the major sources of uncertainties are regarding the range of the sensor, the frequency at which mobile devices generate probe requests, the way and rate at which the mobile devices randomise their MAC addresses, the collisions caused due to the hashing of the MAC addresses, and finally the gaps

introduced by the failure of the sensors. There is an inherent bias to these data caused by mobile phone ownership in the population which varies across time, location and demography. We discuss each of these uncertainties and biases in detail below.

Table 3.5: Summary of the collected datasets.

Dataset	Locations	Time	Detail	Purpose
Small Experiments	3	30 - 60 mins	High	Exploratory analysis
Pilot Study	5	6 weeks*	Medium	Devising and calibrating methodologies
Smart Street Sensor	1000*	4 years*	Low	Real world insights

*approximate

3.5.1 *Range of the sensor*

The first and foremost uncertainty we face with wireless sensors such as Wi-Fi and Bluetooth is the delineation of the field of view of the sensor. Although the Wi-Fi signals can be partly managed or restricted by manipulating their strength, there is no reliable way to precisely delineate a survey area for these sensors. The manipulation of signal strength can be done by installing metal shields around the sensors to block certain directions and prioritise others but the method cannot block out all the signals and will leave some uncertainty about where the probe requests are coming from. Moreover, strength of the signal received from a mobile device by the Wi-Fi access point depends on numerous factors such as,

1. Distance between the mobile device and the Access Point.
2. Thickness of the objects present in-between them.
3. Nature of obstructions, e.g. metal vs glass
4. Interference from other wireless devices.
5. Power level of the transponder of the Access Point.
6. Power level of the transponder of the mobile device.
7. Atmospheric conditions such as humidity, temperature, etc.

The signal strength drops non-linearly when moving away from the Access Point as shown in Figure 3.19 and there is a *close-range non-monotonicity* as well - where within 10 feet, a closer device can report lower signal strength than one further away [Cisco, 2008]. The relationship between the two is given by the equation [5] 3.1,

[5] Zengrzengr, Andy, and Cabral. Calculate distance from rssi, 2017. URL https://bit.ly/2Ohskf9

$$\log_{10} d = \frac{(P_o - F_m - P_r - (10 \times n \times \log_{10} f) + (30 \times n - 32.44))}{10 \times n} \quad (3.1)$$

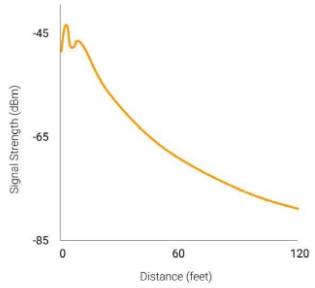


Figure 3.19: The decay of signal strength (RSSI) with respect to distance.

Source: Wi-Fi Location-Based Services, Cisco

Where,

d = distance - Sensitivity of the receiver

F_m = Fade Margin - Sensitivity of the receiver

n = Path-Loss Exponent, ranges from 2.7 to 4.3

P_o = Signal power (dBm) at zero distance - Measured by testing

P_r = Signal power (dBm) at distance - Measured by testing

f = signal frequency in MHz - Specific to the hardware

All these factors vary widely in real world conditions at each location depending on where and how the sensors are installed. They also vary widely over time due to changes in the context, and vary across different directions at each location as well. This makes it extremely difficult to model the distance between mobile device and the Access Point as a function of the received signal strength. The Equation 3.1 can be approximated and simplified as,

$$R = (-10 \times \log_{10} d) + A \tag{3.2}$$

Where R is the reported signal strength and A is the signal strength at 1 metre. Although Equation 3.2 can help us to roughly infer the distance of the mobile device, the uncertainty of this method in respect of location makes it lose the meaning when compared across locations.

From the above we can conclude that it is almost impossible to delineate the field of measurement precisely and accurately by simple methods using the information present in the probe requests. This leads to uncertainty in the data collected which needs to be resolved with explicit assumptions or specific methods to reduce the resulting noise. We require a method to isolate this noise from data generated by devices within the field of measurement. This method needs to be independent of the micro site configuration and temporal changes in the context.

3.5.2 *Probe request frequency*

The second uncertainty we face is that the frequency at which mobile devices generate the probe requests which varies wildly. The number of probe requests generated from a mobile device depends on,

1. Manufacturer of the device. E.g. Samsung vs Apple
2. Version of the software running on the device. E.g iOS 7 vs iOS 8
3. State of the device. E.g. Is it already connected to internet? Has location services been switched off?

4. The number of Access Points already known to the device.

Studies done by Freudiger [2015][6] have shown that the number of probe requests generated by a mobile device vary widely across manufacturers such as Samsung, Apple and LG, across different states the devices are in (such as charging, screen being on, Airplane mode being on, etc.), and depends heavily on the number of Access Points known to the device. It is also seen from our initial experiments that these probe requests are generated in short bursts rather than being generated at regular intervals. This makes predicting a base factor for calculating the number of mobile devices based on the number of probe requests received much more complex. The variety of device models available, and the pace of change in software that run these models, further complicates this. Though we can simply aggregate these probe requests based on the unique information in them, in absence of such information understanding frequency of probe requests becomes extremely critical. We need to consider this uncertainty in detail while making any simple assumptions on the relationship between number of probe requests and the number of mobile devices that generated them.

[6] Julien Freudiger. How talkative is your mobile device?: An experimental study of Wi-Fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '15, pages 8:1–8:6, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3623-9. DOI: 10.1145/2766498.2766517. URL http://doi.acm.org/10.1145/2766498.2766517

3.5.3 *MAC address randomisation*

Randomisation of MAC address is one of the recent uncertainties introduced in the data. As we saw in Section 3.1, the MAC address is the unique identifier for each mobile device and we aggregate the footfall numbers based on this. Since the probe requests are transmitted unencrypted and can be received by any Access Point, this is one of the biggest leaks of personal data which occurs in the Wi-Fi based communications. Modern mobile devices solve this problem by using a randomised MAC address for the probe requests which can result in large over-estimations of the number of mobile devices in the vicinity.

The method of randomisation and the frequency of randomisation varies widely between device manufacturers and also changes as new versions of the software are released. This seriously affects the usefulness of the data long-term, where methods designed to overcome this randomisation can be rendered inefficient in the future. Figure 3.20 shows the increase in the share of randomised MAC addresses since 2015. We can observe that in addition to the overall upward trend there are bursts of increase around late 2016 and 2017, which coincides with the release of new mobile operating systems. This makes it necessary for devising a method to overcome MAC randomisation to be able to

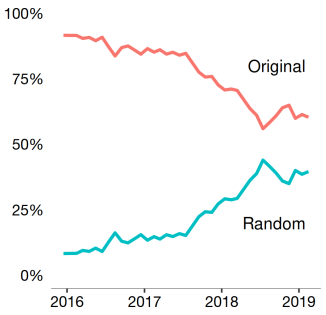


Figure 3.20: Increase in the share of randomised MAC addresses compared to non-randomised original ones over the years. *From data collected at Regent Street, Cambridge.*

uniquely fingerprint devices so that they can be aggregated together. As we saw in our literature review, this is also one of the major opportunities in research on human mobility using Wi-Fi data.

3.5.4 *Mobile Device Ownership*

One of the major external biases in all the datasets collected from mobile devices is the overall volume and nature of the ownership of these devices. The ownership of mobile devices, specifically Wi-Fi enabled ones, have been on the rise since 2005. Although mobile ownership has reached unprecedented levels in recent years, there is still an underlying increasing trend present in the ownership of these devices which manifests itself in the collected data. Moreover, mobile ownership varies widely between demographies of age and geography as well. Figure 3.21 shows the mobile ownership across age groups in the UK from 2012 to 2018. We can observe that the older age groups are under-represented in our data. This needs to be taken into consideration while using this data to extrapolate any demographic conclusions from it. In addition to this, the overall upward trend needs to be adjusted assuming a 1% increase monthly and 0.2% weekly when using this data across long periods of time.

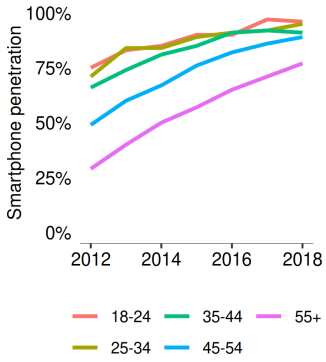


Figure 3.21: Smartphone penetration by age group in United Kingdom. *Source: UK edition, Deloitte Global Mobile Consumer Survey,*

3.5.5 *Missing Data*

Being collected by a distributed set of sensors located in busy real-world scenarios, the data have a large number of gaps as well. These gaps are caused by various reasons such as: failure of the sensor hardware and software, and failure in internet connectivity to send the data back. External factors such as store closures which can cause power loss, regular disruptions such as software updates and maintenance, and finally other factors such as unauthorised tampering and unplugging of the sensor can also cause gaps in the data. This is leads to a dataset which contains several small and medium sized gaps as shown in Figure 4.2. Moreover, the Smart Street Sensor project is implemented and managed with commercial motives: the sensors are installed and uninstalled at locations as retail partners join and leave the project. This leads to an uneven availability of data across locations over longer time periods which creates challenges while aggregating the data across locations. We need to implement a methodology to fill in these gaps which considers the periodic patterns in the data. We also need to devise a measure for aggregating the counts across locations which removes the bias

introduced by long-term gaps in data.

3.5.6 *MAC address collisions*

Finally, from the initial analysis we have observed that there are few instances of collisions occurring in the hashed MAC addresses. This has been observed as unique hashed MAC addresses appear at different locations within a short period of time which cannot be explained by the physical travel by the user between these locations. These collisions are caused by the limitation of the hashing algorithm used and exist only in very large amounts of data. It is important to note that this collisions are specific to non-randomised MAC addresses as we don't expect any consistency within the randomised ones. Even though this is an inevitable side effect of the hashing process, the probability of such occurrence is very low and is calculated as 2^{-n}, where n is the number of bits in the output of the hashing algorithm. The total number of estimated collisions between m unique values is given by, $2^{-n} \times \binom{m}{2}$[7]. This translates to around 100 collisions across a million unique devices with a 32 bit algorithm and 2 collisions across 10 Billion devices when using a 64 bit algorithm. Though these collisions might cause issues in granular mobility models, for long-term and broad studies where we don't track individual devices, they can be safely ignored.

[7] Mikeazo and Poncho. Formula for the number of expected collisions, 2015. URL https://bit.ly/2YS6zYl

4
Processing the Data into Footfall

Chapter 3 detailed the procedure through which three distinct Wi-Fi probes based datasets were collected ranging from small controlled experiments to comprehensive national level project. It also detailed the various uncertainties, biases and challenges that are present in these datasets and possible approaches to solve them. Having established these, this chapter aims to devise the toolkits and methods to deal with the datasets and combine them into a data processing pipeline which can to go through the data and convert them to estimation of ambient population or footfall at the locations they were collected.

As the amount of data collected with Smart street sensor project is large from a conventional computing point of view, Section 4.1 starts with a review of 'big data' and tries to set up a framework of evaluating the datasets from a 'big data' perspective. A brief review was conducted on the topic of 'big data and big data tools' which established a framework for investigating datasets and measuring the extent of 'bigness' in them. Using this framework, the datasets were evaluated in each of their dimensions to understand their nature and the challenges posed in these dimensions. We find that the Wi-Fi based datasets are 'medium data' (as opposed to 'big data') which can benefit from customised toolkits which increases the efficiency. After evaluating the datasets, a detailed review of tools and methods to deal with big data was conducted and the ones which are relevant and feasible for further research were picked out. Finally a complete bespoke 'toolkit' was created by pulling together and connecting all the individual tools so the data can be processed in the most efficient way.

Having designed a toolkit, Section 4.2 explores the methods that can be used by this toolkit to clean and process the data. The major uncertainties in these datasets which were identified in the last chapter were looked in to further with the specific focus on how much they

affect the datasets. We identify - the uncertain field of measurement and MAC randomisation as the biggest sources of data. We discussed the ways in which these problems could be solved and design methods to solved them. We propose signal strength and its analysis as solution for enforcing the field of measurement and sequence numbers and their analysis as solution for figuring out unique devices even when they were randomised. We formalise these ideas into algorithms and use these on the datasets one by one to find which ones are feasible and eliminating the ones that cannot help. Both the methods are tested extensively on the data collected from the initial survey and the pilot studies and the corresponding effectiveness in reducing errors were measured. Finally, an alternative method to adjust long term errors quickly and efficiently in large projects such as Smart street sensors was devised and tested to provide us with footfall estimations of sufficient quality.

Finally in section 4.3 we combine these tools and method together to make a data pipeline which takes in the large amount of the continuous inflow of data from the smart street sensor. This pipeline downloads, cleans, processes and stores the data. It also post-processes the data into footfall for further analysis. The performance and the efficiency of the pipeline is briefly discussed and compared with traditional methods. Thus completing our journey from raw Wi-Fi probe requests data to an informed estimate of footfall at retail location all around UK.

4.1 Data Toolkit

Big data and its analytics promises huge benefits in terms of value realisation, cost reduction and insights but it also introduces a numerous pitfalls [1]. With developments in information technology, mobile communications and the internet of things, large assemblages of data are readily available leading to immense possibilities in research. But when we analyse these data at such scale, we also encounter a large amount of added complexity and cost. Hence it is important to be careful in choosing the methods and tools in dealing with big data where we should look to devise right methods and tools for the right problems. Moreover in several disciplines, such as statistics and geography, the existing methods and tools are already developed for dealing with large scale data. These methods along with improvements in hardware has made the processing big data in these disciplines possible without major changes in workflow. In the current environment of constant change and growth of sources of data, we cannot afford to lose the opportunity to extract information from them while trying to create a perfect, future proof approach in dealing with them. We need to move fast with a pragmatic approach where we look at other disciplines and adopt best practices and solutions in them and develop consistent approach for our needs rather than reinventing the wheel.

[1] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015

In the previous chapters various methods we devised to collect and process data from Wi-Fi probe requests emitted by phones have been discussed in detail. Though we discussed the methods conceptually, we left out the rationale behind choosing the toolkit employed to implement those methods. We start by discussing the concept of 'big data' in general and look at previous literature to understand its definition, nature and the challenges they pose. Then we look at the data-sets we collected through the pilot studies and the 'Smart Street Sensor' project and evaluate them in terms of the dimensions of the big data. We also discuss the challenges faced in dealing with our dataset in detail and try to understand the requirements for devising a toolkit for it. Finally we put together a toolkit to suit our datasets built from simple small UNIX tools. [2]

[2] "Write programs that do one thing and do it well. Write programs to work together. Write programs to handle text streams, because that is a universal interface.", Doug McIlroy on UNIX philosophy.

4.1.1 What is 'big data'?

With the proliferation of internet enabled personal devices, we have quickly moved from data sparse environment to a data rich one. We can even confidently say that we are in an age of data deluge where

the amount of data which are collected and stored are increasing exponentially in a very short period of time [3]. As we saw in the previous chapters collecting large amount of data is quick and easy. Technological advancements have enabled us to be able to think about utilising such large assemblages of data which would have been impossible even in the recent past. By providing unprecedented coverage, these large assemblages of data - 'Big data', provide us with insights which were not possible before. They often change our approach and methods employed in entire disciplines. For example, in computer science, fuelled by the explosion of collected user data, there is a paradigm shift in Artificial Intelligence with the use of data mining, machine learning and deep learning. It is only a matter of time before this approach pervades social sciences research as well. In addition to the above advantages, 'Big data' because of their nature also introduce numerous challenges in their collection, storage, analysis and visualisation. This is not including the enormous additional overhead and complexity introduced when we try to employ big data methods and tools. If we are not careful, using big data tools and methods for solving problems that do not require them can be counterproductive where the advantages realised do not justify the overheads introduced. Hence it is important to understand the 'Big data' nature of the datasets we are dealing with at a granular level and choose the tools and methods without any presumptions.

[3] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1 (1):2053951714528481, 2014

The first and foremost challenge we face while discussing big data is its definition. It is hard to clearly and objectively define 'Big data' as it can vary widely based on the discipline and perspective. What may be 'big' in one discipline may not be in another. The nature of data can also be evaluated in various dimensions and can exhibit different properties in those dimensions. 'Big data' is generally defined within the context of disciplines, as data which cannot be managed with traditional methods and tools in those disciplines and requires substantial change in the approach of the practitioners. This approach of looking at 'Big data' is too subjective and falls short of giving us more understanding of 'Big data'. One of the most quoted definitions pertains to the scale of the data in the dimension of volume - size of the data, velocity - speed of the data and variety - the complexity of the data [4]. This has also been extended to include more dimensions such as, veracity - the reliability or truthfulness of the data, visualisation [5]- the complexity in visual interpretation and presentation of the data, and others such as visibility validity, variability, volatility and value. There have also been other alternative dimensions

[4] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001

[5] Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115:119–133, 2016

proposed such as cardinality, continuity and complexity [6]. However we can consider the core dimensions of data - volume, velocity, variety, veracity and visualisation for evaluating our datasets. Since not all data is 'Big' in all these dimensions, we need to evaluate the 'bigness' of the data in each dimension and consider the associated challenges and solutions.

[6] Shan Suthaharan. Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):70–73, 2014

The second set of challenges arise while we process the big data, its acquisition, storage, extraction, cleaning, annotation, integration, aggregation, modelling, analysis, visualisation and interpretation. Challenges in each one of these processing activity arises due to the data being big in one or more dimensions. The data being big in volume, velocity and variety poses challenges in data acquisition, aggregation, cleaning and analysis. These challenges make traditional methods impractical and introduce the need for distributed, crowdsourced collection of data, heavily parallelised computing and application of functional programming concepts. The unstructured nature of the big data also introduces notable biases which mandate careful consideration, proper calibration and weighting during analysis so that we can understand and remove any uncertainties arising from them. The data being big in veracity dimension poses significant challenges in its analysis and modelling. Since simple methods such as linear regression fails in such scenarios, we require complex methods such as support vector machines, neural networks and hidden Markov models which compensate the lack of structure with the volume of data. With such computationally intensive methods, heavily parallelised high performance computing techniques such as GPU processing become indispensable. We also face significant challenge in visualising such complex features and methods which not only supports critical decision making but also is indispensable in exploratory analysis. The volume and velocity of big data makes them hard to visually simplify and digest. They are especially complex to interpret in the time dimension unless presented in small parts. Geographic information systems do a good job in visualising complex geographic data but struggle to maintain legibility and meaning when dealing with the temporal dimension. The visualisations of big data need to be highly processed, simplified and interactive to present meaning to the viewer. They have to balance between functionality, aesthetics and performance. Finally, because of the variety, big data creates need for consistent, well engineered standards so that multiple approaches and tools can be employed in tandem.

Apart from these processing challenges, we also have management

challenges associated with big data such as privacy and security, data governance and ownership, data and information sharing, and cost[7]. Since these big datasets are usually comprehensive, securing them and protecting the privacy of the users becomes a central consideration in any project dealing with them. In many cases, though the data collected itself may not contain personal information but at these scales, in conjunction with other datasets, it can be used to infer them. The overall approach, methods, tools should comply with relevant legislation such as GDPR as well as the research ethics of all the stakeholders. This is especially challenging since these large unstructured datasets exhibit ambiguity of their ownership as well which calls for a clear, transparent and secure way to share them with other stakeholders along with publications of results in a timely, accessible manner. The associated project management and tracking tools need to be capable of handling these data ownership and sharing concerns as well.

[7] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014

Finally, the biggest challenge we face with big data is the cost in terms of money, resources and time. Though most of the big data tools are developed openly and distributed freely there can be lot of incidental, non-direct costs associated with collecting, processing and managing data with them. For example, there are the operational costs collecting data at such scale, network costs moving them, server costs storing and processing them, cost of procuring and supporting specialised tools and the human resource cost in hiring and training people who are capable for dealing with them. Though there are economies of scale at larger scales, the overall resources required to manage big data effectively can be several folds of what is needed for a traditional dataset. This makes it important to look at the data in our hands closely and carefully so that we can make informed decisions on how 'big' it is and choose the methods which are the most suited for such dataset.

4.1.2 How big are the Wi-Fi probe request datasets?

In this section we take a detailed look at the three sets of Wi-Fi probe requests collected as described in chapter on data collection using the 5Vs big data framework. Our aim is to understand the nature of the data in each dimension and thus evaluate the challenges we face in that specific dimension leading to a bespoke solution. We look at each set of data in each dimension and try to answer the following questions,

1. How can this dimension be measure objectively?
2. How big is the data in terms of the defined measurement?

3. How does it data compare with datasets in other disciplines?
4. How can we describe the size of the data?

We then combine these isolated evaluations to form a combined description of the datasets. This is then used as the basis for developing a list of requirements for designing the data processing and management toolkit.

Volume

Probe requests data, being dynamic and continuous, cannot be quantified as an absolute static number in terms of volume. Hence we use a long term measurement - yearly rate, for each location instead. On shorter datasets such as the pilot study, we estimate the yearly volume linearly from the available data. We standardise this measure as the amount of disk space needed to store the collected data when encoded in text form. It is important to note that this can be reduced many folds by using compression or binary formats but we chose text since it the de-facto standard for exchanging data.

Table 4.1: Comparison of volume or size of the datasets of Wi-Fi probe requests.

Study	Maximum (GB per year)	Minimum (GB per year)	Average (GB per year)	Total* (TB per year)
Pilot Study	134	3	54	48.3
Main Study	6.1	2.4	4.42	4.1
Smart Street Sensor	5.4	0.001	0.8	0.8

** Estimated for 920 locations*

We can see that there is a lot of variability in the volume of probe requests generated at a given location. This mostly depends on how many mobile devices are present around the location. We observe that when we collect most of the information present in the probe requests in a busy area such as Oxford street in the Pilot studies, we generate around 50 terabytes of data in a year. As explained in Chapter 3, in a real world setting such as the Smart Street Sensor project where the sensors fail at times and the amount of data collected is optimised, the volume is around a 1 gigabyte. The total volume of data we deal with in the case of a national scale project with around 920 sensors running for around 4 years is around 2 terabytes. A comparison of this to datasets from other disciplines is shown in Figure 4.1. It is key to note that the y-axis is scaled exponentially.

We can see that the probe requests data is not truly 'Big data' as experienced in other fields. It is only when we reach a complete coverage, i.e, putting a sensor at each retail establishment in UK, our estimated

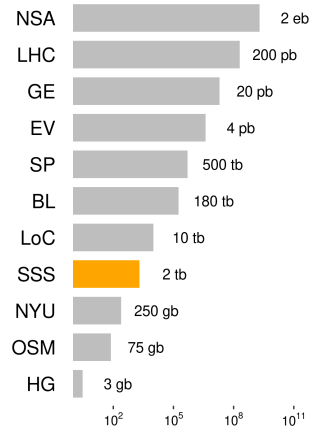


Figure 4.1: Comparison of volumes of data across various disciplines.

NSA - National Security Agency, LHC - Large Hadron Collider, GE - Google Earth, EV - Event Horizon project, SP - Spotify music, BL - British Library data store, LoC - Library of Congress, SSS - Smart Street Sensor, NYU - New york city Uber trips 2009-15, OSM - Open Street Map and HG - Human Genome Project

data volume reaches around 250 petabytes which is comparable to scales experienced in other fields such as particle physics and world wide social networks. At the same time, the scale of probe request data is not small either. The volume of 2 terabytes is more than the memory available in any desktop systems and is more than any of them can process in a timely manner. Summarising from the above, we can confidently say that the probe request datasets are 'Medium Data' - especially the dataset collected by the smart street sensor project. Though it has potential to scale into a truly big dataset, for the purposes of this research we can classify it as 'Medium data' in the volume dimension.

Velocity

Velocity is the rate at which the data is collected over time. It is significant when evaluating big data since some data which may not scale in terms of absolute volume but the speed at which they are received makes them challenging to deal with. A perfect example is the comparison between data generated by the Large Hadron Collider project by European Council for Nuclear Research and a world wide social network such as Facebook. Though their total volumes are comparable at 200 petabytes, the data from LHC is generated in concentrated experiments at a rate of 3 petabytes in 5 seconds while Facebook generates the same about in about a day or two. Since the size of an individual Wi-Fi probe request doesn't vary widely, we define the velocity of this dataset as the number of requests received at a given location at a given location within a given time interval. Though the precision of the time measured during data collection is in microseconds, the practical data transfer resolution in all the datasets is around 5 minutes. Hence we measure velocity of out datasets in terms of number of requests every 5 minutes. Table 4.2 compares the datasets we collected on Wi-Fi probe requests in terms of their volume.

Table 4.2: Comparison of velocity or speed of the datasets of Wi-Fi probe requests.

Study	Maximum (per 5min)	Minimum (per 5min)	Average (per 5min)	Total* (Mn per 5min)
Pilot Study	8577	188	3469	3.20
Main Study	1362	534	782	0.72
Smart Street Sensor	5024	6	408	0.27

** Estimated for 920 locations*

We observe that locations can receive up to 8500 requests in 5 minutes or can get no request at all depending on the time and how busy it is. But we can see that on average a national scale project with around 900 locations generates around a million requests every 15 mins. Com-

pared to the LHC's 180 billion records or Google's 190 million searches per 5 minutes this seems to be not high speed data. However, this is much faster compared to traditional data sources such as census or geographical surveys which are updated anywhere between 6 months to 10 years.

To summarise, in terms of velocity, the Wi-Fi probes data can be described as 'Medium' at best. The methods dealing with the data should be time sensitive and be able to deal with a continuous stream of data but at the same time need not be real time or need sub-second latency. Since the Wi-Fi probe requests don't have actual location information the mobile devices, it does not have the similar value in real-time analytics as shown in comparable location or movement based datasets.

Variety

Variety is defined by the amount of variance in the type and characteristics of the data. Since variety is hard to quantify and compare across disciplines we evaluate the dataset subjectively for the variety present in it. The data transmitted in a Wi-Fi probe request is defined by the 802.11 Wi-Fi specification [8] and every probe request has to have a set of mandatory fields for Wi-Fi to work. This set of fields is the same everywhere across the world and the specification, especially the probe request part, has remained stable over years. Though there is some variability allowed within the specification, being part of a global standard, the data collected is heavily structured in general.

[8] IEEE. IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, Dec 2016

The first set of variety present in the Wi-Fi probes data set arises from the 'information elements' part of the probe request. The structure of a probe request is discussed in detail in the data collection chapter and is summarised in Figure 3.1. Essentially the information about the capabilities and type of the mobile device is encoded in the information elements part of the probe request and this information is optional and is implemented at the discretion of the manufacturers. As this information elements are demonstrated to be useful in successfully fingerprinting the mobile devices [9], mobile devices increasingly don't include any information in them. Emergence of manufacturers with large market share and narrow set of device models such as Apple and Samsung also reduce further variability in them. The second set of variety in the dataset arises from the rate at which these probe requests are generated by the mobile devices. Unlike devices which generate data on events or at regular intervals, mobile phones generate probe requests at a rate based

[9] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, Frank Piessens, and Piessens Frank. Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pages 413–424. ACM, 2016. ISBN 1450342337. DOI: 10.1145/2897845.2897883

on various factors. Though this leads to some challenges in counting footfall from these probe requests the variability exhibited here is neither so large nor so complex that traditional methods could not deal with them.

Comparing with some of the big data encountered in unstructured data collected over web such as social networks or other sensor based methods, the variability here can be considered trivial. Further when we convert these probe requests in to footfall counts, the variety in the dataset drops almost to zero as it becomes just an ordinal data point varying in geography and time. Summarising the above, we can confidently say that the Wi-Fi probe request data does not exhibit any 'big data' properties in the variety dimension.

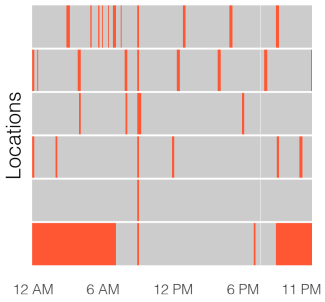


Figure 4.2: Missing data from five locations at Tottenham Court Road, London on 15 January 2018 demonstrating the veracity of the data.

Veracity

Veracity is defined as the amount of abnormality present in the data in the form of inaccuracies, biases and noise. Similar to variety, veracity is hard to quantify hence required a subjective evaluation. Being sensor collected data, veracity is the dimension where the data exhibits most 'big data' properties.

First set of veracity in the dataset arise from the fact that it is collected through sensors located in multiple locations which communicate to the central server using 3G mobile data connectivity. We know from experience that the sensors are unreliable and fail to send back data regularly due to various reasons. More over the sensors are installed and uninstalled regularly as partners join and leave the project. This results in a data stream which is often erratic and incomplete with large gaps in them. In addition to this the sensors need to be rebooted regularly due to issues or updates leading to small gaps as well. Since the sensors are part of retail establishments they can be switched on and off regularly in some of them as well. Figure 4.2 demonstrates the veracity of the data in terms of missing data for a sample of 5 locations in London. All the above pose immense challenges when we attempt to aggregate the data where we have to estimate and fill these gaps.

There is also a lot of variability in the physical location of the sensors and the area of measurement. The sensors may report higher or lower count due to their configuration and the context of their location as discussed in chapters pertaining to data cleaning. This leads to a situation where the accuracy of the data collection varying quite widely across location and times [10]. It is often not clear if the change in the data is due

[10] Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of GISRUK 2017*. GIS Research UK (GISRUK), 2017

to actual changes at the location or just the change in the configuration of the device. For example, opening of a mobile shop next door to the sensor can increase the estimated footfall without any change in actual footfall at the location.

Finally we also have to work within the changing mobile landscape. Though the Wi-Fi probe requests are standardised by IEEE, the mobile manufacturers have started adopting obfuscation techniques to protect the privacy of the users. This started with randomisation of MAC addresses, removal of information elements and generally getting more sophisticated with new versions of operating system. There is also bias in terms of operating system adoption and change in market share between manufacturers. There is no inherent structure or information on what is changed and how often these changes occur which leads to questions on the continuity of the data over long periods of time.

Summarising from the above, we can confidently conclude that Wi-Fi probe requests dataset shows 'Big data' characteristics in terms of its veracity and requires appropriate tools and methods when aggregating, analysing and modelling it.

Visualisation

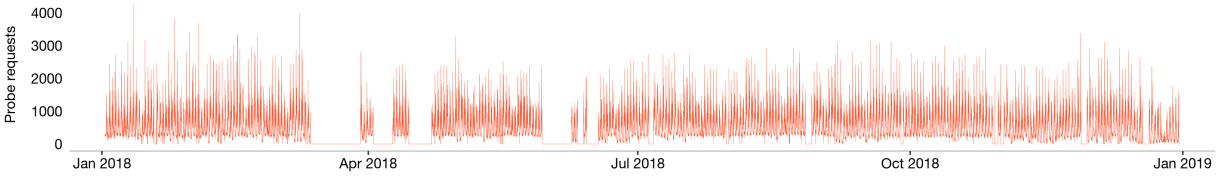


Figure 4.3: Number of probe requests collected for every five minute interval at Tottenham Court Road, London on the year 2018 showing the visual complexity of data in the time dimension.

Visualisation is closely related to volume, velocity and variety of the data. The Wi-Fi data due to its non-trivial volume and velocity, exhibits similar characteristics and challenges in terms of visualisation. Since there is not much variety in the dataset, when we process the raw data into footfall counts we are left with just the time, location and footfall count for each data point. Out of these, location and footfall counts are easy to visualise but time exhibits big data properties. This is primarily due to its granularity at 5 minute intervals and longitudinal nature of the data collection. The major challenge with Wi-Fi data is to simplify and visualise them in a legible way while showing change in term of time. The veracity of the data presents challenges in simplifying them and the volume poses challenges in maintaining legibility. We also have

to take the 'near real time' aspect of the data into consideration while visualising them. There is a clear need for always on, interactive, real time dashboards with geographic capabilities in addition to the capabilities of traditional desktop GIS. There is also need for multiple linked dynamic visualisation platform for separating the scope of the visualisation into manageable units. Figure 4.3 demonstrates the illegibility of simple visualisations of the data due to granularity, variability and veracity. We can safely say that the Wi-Fi probe requests dataset is at best 'Medium' in the visualisation dimension.

Summarising the above discussion, we can conclude that the datasets collected from Wi-Fi probe requests are at best of 'medium'. They show the most big data characteristics in terms of their veracity. In rest of the dimensions the datasets are not truly big data and we need to look at tools and methods appropriate to their size. The toolkit we devise need to be able to deal with their mid-size volume, velocity and visualisation dimensions and at the same time need to able to deal with the large amount of veracity of in them. Figure 4.4 illustrates the summary our discussion. This leads us to devise a 'medium data toolkit' which can be used without incurring the extra cost and complexity introduced by big data tools while be able to handle the data at hand.

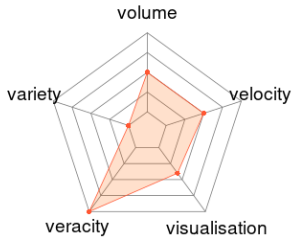


Figure 4.4: Big data characteristics of the Wi-Fi probe request datasets in their corresponding dimensions

4.1.3 *A Survey of Methods and Tools*

Having classified the Wi-Fi probes dataset as a 'Medium' sized data, in this section, we survey the tools and methods available at various stages of the data processing and management process - data collection, storage and retrieval, processing and analysis, visualisation. We first survey the tools available in each stage and specifically look at their suitability for Wi-Fi probe request datasets in terms of the following characteristics,

- *Performance* - How much data can be processed in a given time?
- *Flexibility* - How easy it is to change the scale and scope?
- *Complexity* - How many components or parts are involved?
- *Cost* - How much money or infrastructure do they require?

We then discuss the principles of UNIX philosophy and how it helps in solving similar sized problems in computer science. Finally we pick and connect the tools to devise our toolkit which is best suited for out Wi-Fi probe request dataset.

Collection

We discussed various technologies used in collecting passive data on ambient population and pedestrian movement in the literature search. In this section we look at tools and methods used to collect Wi-Fi based data passively. The primary considerations for evaluating data collection strategy are the scale of the infrastructure, expertise and effort required to implement it and cost involved.

There have been numerous sensors, tools and associated software platforms made available for data collection under the umbrella of 'internet of things'. We start by looking at different approaches in the Wi-Fi data collection tools and try to reason the most appropriate solution for our research. On one end there are low level and low cost bespoke solutions which require lot of effort to implement and maintain. On the other end there are turn key solutions which doesn't require lesser effort but costs considerably more. The key is finding a balance between both while satisfying the requirements of the project. Since the Wi-Fi data is medium sized in terms of volume and velocity, we can deal with solutions with less than optimal scalability but since the data is 'big' in terms of veracity the toolkit has to give us most flexibility. Essentially, we are looking for a data collection methodology which prioritises flexibility and cost while performing moderately in terms of scalability and complexity as illustrated in Figure 4.5.

Type of solution	Examples
Bespoke	Micro-controllers with Wi-Fi modules e.g. Audrino + ESP8266
Turn-key	End to end commercial services e.g. Blix, Euclid, Pygmalios etc.
Ideal	**General purpose hardware e.g. Raspberry Pi, Repurposed mobile devices - Tablets, Phones etc.**

Table 4.3: Examples of different types of Wi-Fi based data collection solutions.

In terms of hardware, an example of a highly customised solution would be a micro-controller, such as Arduino, coupled with dedicated Wi-Fi module and programmed with custom software to collect the exact data needed. Designing and implementing of such system is time consuming, cumbersome and usually involves significant cost but it can also be highly flexible, efficient and cheap to deploy. On the other end of this spectrum, we have end-to-end solutions such as Blix, Walkbase, Euclid, Retail Next, Pygmalios, etc. where the data is collected through multiple sensors and sources and syndicated into a clean footfall information by a third party service provider. These platforms for footfall data collection and analysis have the advantage of being quick and easy

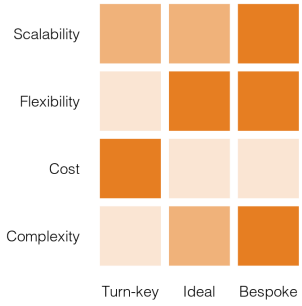


Figure 4.5: Characteristics of types of Wi-Fi data collection tools at each end of the spectrum compared to an ideal candidate

(Darker colors show higher score)

to develop and deploy while they can also be highly inflexible for changes and turn out to be costly when scaled up. A middle ground here is to use a general purpose hardware such as single board computers or repurposed mobile devices, augment them with additional hardware modules and use general purpose scripting languages to write software for them. This way we avoid low level hardware or software design and implementation while maintaining good amount of flexibility. Table 4.3 shows some examples of such systems while highlighting an ideal system.

The Smart Street Sensor project uses its own proprietary sensor system designed and instrumented by the data partner. The design and implementation decisions were made with the commercial application in mind and is not entirely relevant to our discussion in the context of our research. For the research conducted with the data, it is necessary to understand the data collection process and make sure it aligns and integrates with the rest of the toolkit. As discussed in the data collection chapter, the methodology used in the smart street sensor project satisfies our requirements. The toolkit we designed to collect other datasets are in-part inspired by this methodology or a modified version to include more flexibility. The toolkit consists of Raspberry Pi, Linux, tcpdump or tshark [11] and nodejs. Raspberry Pi and the Linux OS provides a general purpose base system and hence the flexibility. On top of this we built our data collection system by assembling open source and free network analysis tools such as tcpdump and tshark along with other tools providing functions such as scheduling, personal data obfuscation and data transmission with scripting languages like nodejs and bash.

[11] Gerald Combs and Contributors. Wireshark - network protocol analyzer. https://www.wireshark.org/about.html, 2018

Storage

Data storage technology is one of the most diverse landscape in terms of both methods and tools available. It has been constantly in research and development since the beginning of computing and is one of the fastest changing landscapes with the advent of big data paradigm. A comprehensive review of storage solution warrants a chapter in itself so we restrict our survey to an outline of most significant approaches and corresponding systems and tools.

At one end of the spectrum is one of the most underappreciated for data storage - File systems. Though they seem like a low level interface for storing data, file systems have their advantages as well. When the data is not complex or inter-related, flat text files in file systems could

be the fastest way to store, search and retrieve data. Since operating systems are usually optimised to manage storage media through file systems, they involve no additional overhead and are extremely reliable. The hierarchical file systems use in most of the operating systems act as an index with hierarchical data. The major disadvantage of file systems is that they are not useful for managing data with any kind of complexity. This is the primary reason why database management systems are developed on top of file systems.

Database systems can be broadly divided into relational and document based. The relational databases are optimised to deal with relational data and usually enforce strict structure for the data In general they can handle large number of rows and are designed to scale vertically through addition of more resources to the DBMS such as CPU and Memory. Most relational database systems try to guarantee ACID [12] compliance and hence used in critical systems such as financial operations, sales etc [Haerder and Reuter, 1983]. The document based databases are optimised to deal with unstructured data and can doesn't need a strictly defined scheme. In general they can handle large number of columns and are designed to be distributed and scaled horizontally by adding more instances of the databases which balance the load and redundancy between them. Being distributed, most document based databases try to pick a focus and compromise on others as specified in CAP theorem [13]. There a numerous databases systems which prioritise different things and the right solution depends on the properties of the data and the requirements of the project.

Since the publication of the paper on 'Google file system' by Google [14]. There have been significant effort in designing and building 'big data' file storage systems which can large data in the range of petabytes. These systems are designed to be distributed and optimised for high throughput for queries on them. Hadoop Distributed File System (HDFS) is one such file system which is also the most widely adopted. There are numerous cloud based, third-party solutions built with these file systems making them easy to use. There are also numerous tools, libraries and frameworks which emulate the features of database systems on these distributed file systems making them easier to use further. The primary advantage of these systems is the sheer scalability they provide when it comes to data volume. The primary disadvantage is the associated overheads in terms of cost and time incurred in learning, designing and implementing them. Unless the project is sufficiently large, the

[12] Atomicity, Consistency, Isolation and Durability are properties which make sure that the data in the database is valid even during failures.

[13] Brewer's theorem or CAP theorem states that it is impossible to simultaneously guarantee consistency, availability and partition tolerance in a distributed data store.

[14] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, pages 20–43, Bolton Landing, NY, 2003

advantages gained usually do not justify the overheads introduced. Table 4.4 summarises the above discussion along with relevant examples.

Table 4.4: Various data storage approaches and their characteristics.

Approach	Data size	Examples	Comments
File system	< 10 TB	ext, ntfs, zfs, btrfs	Simple and efficient. Best for hierarchical data. Cannot handle complex connected data.
Relational DB	< 5 TB	MySQL, PostgreSQL	Handles structured and relational data. Optimised for large amount of rows and tries to guarantee validity.
Document DB	<10 TB	MongoDB, Cassandra	Handles unstructured data. Optimised for large number of fields and distribution to multiple clusters. Tries to focus on any two guarantees of the Brewer's theorem.
Distributed FS	> 10 TB	HDFS, Ceph, GFS	Optimised for really large datasets which need to be distributed over multiple nodes.
Cloud Storage	> 10 TB	AWS, Azure, SWIFT	Implements distributed file systems on the cloud. Has more reliability and scalability than local implementations.
Data Warehouse	> 10 TB	Hive, Hbase, Impala, Presto	Interfaces built on top of distributed file systems to emulate capabilities of relational databases on them.

We saw that the Wi-Fi probe request datasets are 'Medium' sized hence we can safely eliminate distributed file systems for storing them. Though the smart street Sensor project uses Azure Blob Storage, when the data is downloaded to the local servers at the university, we can just store them in the file system because of their size (2TB approx.) and the hierarchical nature. The folder structure of - year/month/day/location/interval/ with individual text file, enable us to query the data for any given location at any interval nearly instantaneously without any further database operations. When this raw data is processed into 5 minute counts, we require more relational queries. For this purpose a relational database is sufficient as volume is quite small (25GB approx.). We chose PostgreSQL because of the PostGIS extension which gives us flexibility in handling geographic data.

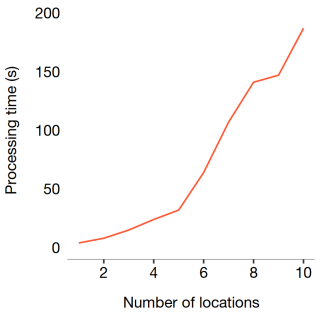


Figure 4.6: Exponential increase in the processing time when using traditional methods.

The processing involves parsing JSON data received for a single day at each location and aggregating them as number of probes requests received in every five minute intervals.

Processing

We saw that the data is medium in terms of volume and velocity and shows big data properties in terms of veracity. Hence we require tools which are capable of dealing with the veracity of the data while being able to manage the volume and velocity. The traditional approach to deal with such dataset is to load it into a general purpose analysis tool

such as R or a GIS package and process it. The size of the dataset and the lack of meaningful complexity of geography element in the data eliminates the use of GIS packages. Scripting languages such as R and Python can deal with the dataset and its requirements but the time taken to do so increases exponentially with the size of the data as the size of objects in memory increases. Figure 4.6 illustrates the increase in processing time with respect to number of location for a simple exercise where a day's worth of raw data is parsed and aggregated into number of probe requests per 5 minute intervals (The code used to produce these benchmarks are detailed at Section 7.4). This becomes prohibitively expensive as the number of locations and complexity of the processing increases. Though this can be improved with more efficient coding practices, the margin of improvement is quite limited hence creating the need for better techniques. It is important to note that data processing is done in two stages - the first stage where the raw Wi-Fi probe requests are filtered, cleaned and aggregated into footfall counts and second stage where the footfall counts are in turn analysed to produce reports and visualisations. The traditional methods are sufficient for the second stage of the processing and the first stage is the one which requires a better solution.

On the other end we have big data analysis tools which are built for dealing with extremely large amount of data. Since the publication of the paper on MapReduce [15], there have been immense developments in the Big data analysis landscape. There numerous distributed programming tools to use the data stored within a distributed storage system each focussing on specific type of data and analysis. A concise, non-comprehensive list of types of data or specialities and corresponding big data tools is shown in Table 4.5.

[15] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1): 107–113, 2008

Tools	Speciality
General purpose	MapReduce, Spark
Real-time streams	Flink, Pulsar
Events or messages data	Storm, Kafka, Flume
Networked or graph data	Tinkerpop, Corona
Scheduling	Oozie, Falcon, Azkaban
Turn-key platforms	SpringXD, Cask Data

Table 4.5: Various types of big data processing tools and corresponding examples.

We can rule out the necessity of the above big data tools since our dataset is neither big enough nor fast enough. The dataset does not have any specialised structure such as graph or network but just a very minimal component of geography to it. Using any of specialised big data tools is just going to introduce immense overheads without any added

benefits. We need something in-between the above two approaches where we is sufficiently fast and flexible for our datasets.

This is where we come across the possibility of using standard Unix tools along with connecting them to create a processing pipeline. In some cases, a data processing pipeline made using command line Unix tools have been demonstrated to be 230 times faster than using big data toolkits [16]. The command line tools were developed as parts of Unix operating system for processing text. They are developed in line with the Unix philosophy which focuses on modular and minimal software development. The core tenants of the Unix philosophy has been summarised by Doug McIlroy as below,[17],

[16] Adam Drake. Command-line tools can be 235x faster than your hadoop cluster, Jan 2014. URL https://bit.ly/2s2XZYI

[17] Malcolm D McIlroy, Elliot N Pinson, and Berkley A Tague. Unix time-sharing system: Foreword. *Bell System Technical Journal*, 57(6):1899–1904, 1978

Table 4.6: Tasks in the processing pipeline, corresponding R libraries and equivalent Unix tools

Tools	R Library	Unix tool(s)
Move data to and from Azure blog storage, SQL server and Postgres	AzureR, odbc , RPostgreSQL	azcopy, mssql, psql
Convert data from JSON format to CSV	jsonlite	jq
Encrypt raw data for secure storage	Rcrypt	gnupg
Anonymise personal data into cryptographic hash	digest	openssl
Transform and manipulate tabular data	dplyr	find, cat, cut, grep, sed, awk, sort, uniq, column, paste, join
Impute missing value using time series analysis	imputeTS	Rscript
Visualise the results into maps and charts	ggplot2	Rscript
Create and manipulate geographic data	sf, rgdal	postgis, gdal

1. Make each program do one thing well. To do a new job, build afresh rather than complicate old programs by adding new "features".

2. Expect the output of every program to become the input to another, as yet unknown, program. Don't clutter output with extraneous information. Avoid stringently columnar or binary input formats. Don't insist on interactive input.

3. Design and build software, even operating systems, to be tried early, ideally within weeks. Don't hesitate to throw away the clumsy parts and rebuild them.

4. Use tools in preference to unskilled help to lighten a programming task, even if you have to detour to build the tools

and expect to throw some of them out after you've finished using them.

These principles along with the 'pipe' operator gives us necessary tools to build more complex tools. We can replace most of the libraries we used in the R implementation of our processing with a corresponding command line tools and connect them together with a text interface to achieve similar pipeline. The first advantage of such design is that it is much more efficient than a monolith design. These tools being actively developed for since their invention are compiled as native binaries and are usually extremely optimised resulting in a much faster pipeline. Because of the design of the pipe operator, the individual parts of the pipeline are executed in parallel as chunks of data are passed through them thus avoiding the need to load entire datasets into memory which results in an exponential increase processing time with the size of the data. Being modular, we can even introduce process level parallelism to parts of the pipeline without any major change in the overall design. Finally the modular structure also gives us the advantage of using the best tool for any part of the pipeline.

All of this gives us an extremely minimal and efficient toolkit to process the raw Wi-Fi probes data into counts in a scalable way. Figure 4.7 compares the processing times of such Unix toolkit with the traditional R based toolkit as the data size increases. We can see that Unix toolkit performs extremely well and the performance gains are significant as the size of the data increases. For example, to process data for 25 locations, R based toolkit takes around 20 minutes while the Unix toolkit gets it done in 20 seconds (The code used to produce these benchmarks are detailed at Section 7.4). Table 4.6 shows the activities in our pipeline and corresponding libraries in the traditional R workflow along with the equivalent Unix tools. It is important to note that tools for doing specialised actions such as statistical analysis, machine learning and time-series analysis are built on top of scripting languages such as R and Python.

These can be embedded into our Unix pipeline as scripts running in corresponding front-ends such as Rscript or python. This toolkit can be further accelerated by parallelising parts of the pipeline using gnu-parallel [18]. For example, the previous example pipeline can be parallelised by spawning a pipeline for each location this reduces the processing time for a set of 25 locations from 18 seconds to 3 seconds. This done by utilising every processor cores available in the CPU. Figure

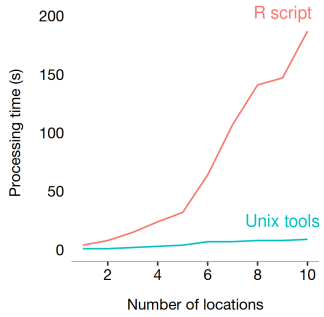


Figure 4.7: The increase in processing time with the Unix pipeline is linear thus improves the scalability compared to R based processing

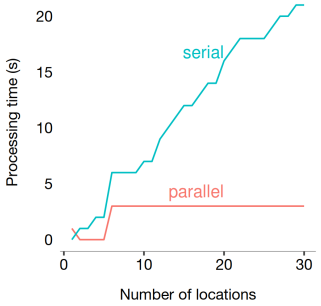


Figure 4.8: The scalability of the processing pipeline could be further improved with parallelising it.

[18] Ole Tange. *GNU Parallel 2018*. Ole Tange, March 2018. ISBN 9781387509881. DOI: 10.5281/zenodo.1146014. URL https://doi.org/10.5281/zenodo.1146014

4.8 compares the processing times of the Unix toolkit with a parallelised implementation (The code used to produce these benchmarks are detailed at Section 7.4). Finally all the Unix tools discussed in this toolkit are open source and free software which has almost no cost in terms of resources. Since these tools are part of the POSIX specification [19] for operating systems, the expertise in their design and use are transferable to and from other disciplines thus reducing researcher time learning and using these tools.

[19] Stephen R. Walli. The posix family of standards. *StandardView*, 3(1):11–17, March 1995. ISSN 1067-9936. DOI: 10.1145/210308.210315. URL http://doi.acm.org/10.1145/210308.210315

Visualisation

In the last section we saw that the visualisation dimension of the data shows some level of complexity. The primary source of this complexity arises from the longitudinal nature of the data and the noise due to granularity of the data. For the processed dataset, traditional visualisation and mapping libraries with R is sufficient while the visualisation of raw data across long time periods for either for exploratory analysis or for communication needs some form of interactivity or simplification to be able to legible. Data driven documents (D3) [20] and Dimensional charting (DC) provides us with both of these requirements. Both of these tools can accept text based input and can fit with other Unix tools discussed earlier. In case of binary file output such as images or documents, they could be directed to the file system and then read into other programs.

[20] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. URL http://vis.stanford.edu/papers/d3

4.1.4 *The Bespoke 'Medium data toolkit'*

In this chapter we saw how the advent of internet and internet enabled devices has lead to significant increase in the amount of data generated and collected across disciplines. This data deluge and improvements in the capabilities of computing hardware has fuelled an explosion of research and development in tools and methods to deal with these 'Big data'. Though these big data tools promise huge improvements in processing capabilities, when used under wrong circumstances they can lead to unwanted overheads and costs. Thus we need a framework to examine and understand the scale of the data that is being used so that we use the right tools for the right purposes and the 5Vs of 'Big data' - Volume, Velocity, Veracity, Variety and Visualisation provides us with such frame work. Every dimension of big data poses unique set of challenges and we need make right decisions in choosing specialised tools and methods to overcome them.

We then closely examined the Wi-Fi probes data we collected with this

framework and found that the data, though posed significant challenges with traditional data processing techniques, do not exhibit 'big data' properties in all its dimensions. Only veracity of the data was found to have any meaningful big data properties, while volume and velocity was found to be 'medium' at best. The datasets lacked any variety and posed minimal challenge in the visualisation dimension because of it high temporal granularity. Thus we arrived at the requirements for a bespoke 'medium data toolkit' which is able to deal with these challenges.

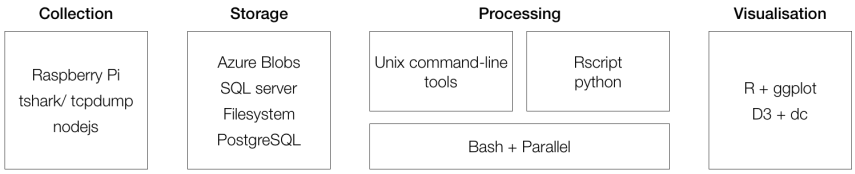


Figure 4.9: Outline of the 'Medium data toolkit' devised to collect, process, visualise and manage the Wi-Fi probe requests data

We undertook a brief survey of tools available for collecting, storing, processing and visualising the Wi-Fi probe request data and with the understanding of the data from the previous analysis chose the ones which are perfect for the datasets. For collecting Wi-Fi probes data in a scalable way, we chose a general purpose single board computers such as Raspberry-Pi along with open source tools such as tcpdump and tshark in a Linux environment. For data storage we narrowed in on using just the file system for the raw data and relational database management system for the processed counts. To process the raw data we chose to devise a processing pipeline using an assortment of standard Unix command line tools linked together using a shell scripting language and parallelised at the process level with gnu-parallel. We also demonstrated that this processing pipeline can be 400 times faster (20 minutes to 3seconds) than the using a monolithic pipeline even with a small sample of locations. For visualisation we chose D3 and DC as the solution for communicating time information legibly. Finally we arrive a 'medium data toolkit', illustrated in Figure 4.9, which is best suited for the Wi-Fi probes dataset which we employ to process and examine the data further.

4.2 *Data processing*

The sources of uncertainties in the Wi-Fi probe request data were discussed in detail in section 3.5. The major uncertainties that were identified were: range of the sensor, probing frequency of the mobile devices, randomisation and hashing of MAC addresses, changing mobile ownership, and missing data. It was determined that the MAC address collision occurring due to the hashing process was insignificant and could be ignored safely. The missing data problem was found to be an issue of 'post processing', whereby the values need to estimated based on the previously occurring values after the rest of the uncertainties were solved. This left MAC randomisation and the uncertainty regarding the range of the sensors as the major sources of noise in the data. The next step was to explore the extent of the noise generated by looking at both the sample data collected and the real world data from Smart Street Sensor project.

In 2015, during the early stages of the Smart Street Sensor project, a data quality audit was conducted to estimate the amount of noise in the data before proceeding to analyse them [21]. A field survey was conducted at locations in Sheffield and London over 5 days in September and December 2016 and manual pedestrian counts were collected for these locations to be compared to the counts reported by the sensors. It was intuitively expected that there would be errors in the sensor collected data, arising from various internal and external factors which may lead to the under- and over-counting at the chosen locations. The MAC randomisation introduced in iOS devices was also expected to be manageable with a standard adjustment factor: a combined measure of ratios of the Apple devices which we observed in sensors prior to the randomisations. Surprisingly, the study found that even with fairly complicated adjustment factors the errors were large and unpredictable and emphasised our lack of detailed understanding of the probe request process. The results showed that the errors varied between 16% to -41% within the same day and although an adjustment factor based on manual data collection reduced the errors to less than 5% there was still a risk of substantially under- or over-counting depending on the quality of the field survey. The study also called for a more closer look at the randomisation process, as well as the development of more advanced data mining methods so as to reduce the reliance on manual counting on the field leading.

[21] Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of GISRUK 2017*. GIS Research UK (GISRUK), 2017

From the initial experiment, the noise due to the uncertain range of sensors was found to be much larger than expected. It was observed that about 53% of the total probe requests collected were from outside the desired field of study. As such the errors could be enormously reduced by simply filtering out the noise It was observed that the signal strength of the probe requests imparts valuable information on whether the probe request is generated by a device within the field of study or not. Moreover, the amount of noise generated was found to be much greater in the data collected in a real world setting at UCL. The experiment showed that the Mean Average Percentage Error when comparing the sensor counts to manual counts of pedestrians, was reduced from 736% to 80% just by removing signal strengths which were lower than -70dBm. However, the problem with that methodology is the arbitrary nature of the threshold -70dBm which can vary widely based on the site conditions and over time. A robust method was needed to calculate this threshold dynamically for each location at a particular time, and which was derived from the data itself rather than requiring external sources of information such as regular field surveys.

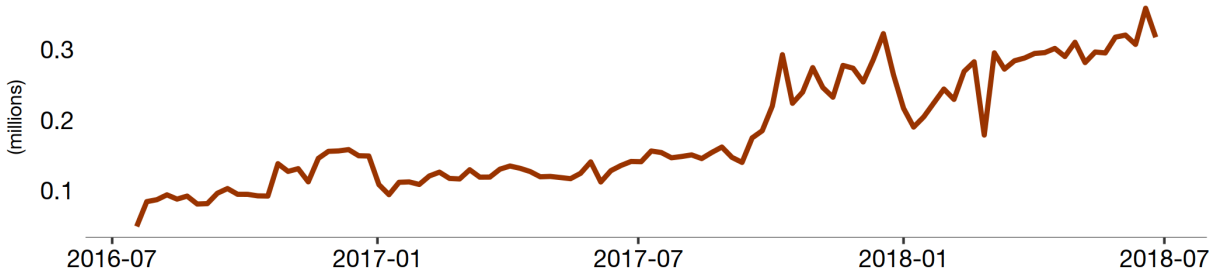


Figure 4.10: The long term effect of MAC randomisation on average weekly footfall estimated at sensors in Cardiff.

As discussed in Chapters 2 and 3 MAC randomisation is a major issue in the data. It has arisen as a result of efforts to protect the privacy of the users of mobile devices, and it cannot be eliminated by simply collecting more data. The belief that we should protect the privacy of mobile device users, also eliminates most of the intrusive techniques for collecting and processing data. The technique of randomising MAC addresses has been used by the Windows Operating System on laptop computers for a long time, but it was brought mainstream when Apple Inc. introduced iOS 8 for their mobile devices in the fall of the 2015. It was observed that there was a massive wave of adoption shortly after the release, but the trend stabilised as more and more of the market share

was captured by Android devices, which were not randomising MAC addresses at the time. The trend remained stable until the autumn of 2017, after which Android devices switched to randomisation techniques. It is important to note that the randomisation methods were neither fool proof nor standardised, as showed by [Vanhoef et al., 2016]. The problem of randomisation continued to intensify in 2017, and was attributed to the change in frequency that MAC addresses were randomised in a given time interval, thus increasing the proportion of local addresses in the dataset along with the total number of unique addresses.

Figure 4.10 shows the resulting 'explosion' in the average number of unique MAC addresses that occurred in September 2017 from a subset of data comprising of sensors in Cardiff. It should be noted that the overall increase in the unique MAC addresses is not due to an increase of footfall at these locations.

In addition to causing problems generally in the data for longitudinal analysis, randomisation also causes issues at specific locations which have the potential for large amount of devices to dwell around them. For example, seating areas in cafes / restaurants, bus stops, and phone shops around the sensors all cause huge overestimations of footfall when aggregated by unique MAC addresses. It was therefore imperative that the method we devised should take both of these cases into consideration.

4.2.1 Methodology

Keeping the above considerations in mind, two methods to clean the Wi-Fi data and process them into footfall data were devised. The first method uses signal strengths to filter out the noise originating from outside the field of view, and the second uses sequence numbers to group probe requests together instead of MAC addresses.

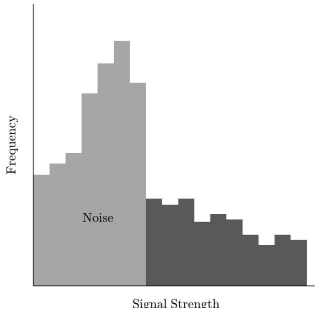


Figure 4.11: Thematic diagram showing the idea behind filtering using signal strength distribution.

Filtering with Signal Strength

One of the clues that we used to estimate the distance between the mobile device and the sensor was the strength of the signal received by the sensor. The first and obvious way to approach this problem was to try and establish a relationship between signal strength and distance. Once established, the relationship was used to convert the measurement to distance, to set a distance threshold for every location, and finally to filter out the probe requests which were outside the distance threshold. However as explained in section 3.5, this approach was found to be unfeasible. The decay of signal strength with respect to distance is not

always constant or linear and is fairly complex to model. Moreover, the parameters with which these two were modelled together such as atmospheric conditions, the presence of obstructions between the source and the target, the material of these obstructions, and the strength of the signal (power level) of the source, vary widely across locations and across time as well. This severely limits the ability to establish a simple conversion factor between reported signal strength and distance. As such, a method which takes in to account all of these variables across the various locations needed to be devised.

Assuming that there are specific patterns in the way a sensor is installed at a location, it was expected that the data from around the sensor should reflect those patterns. That is, in configurations where a specific source of background noise was at a constant distance, there should be a distinct pattern in the number of probe requests reporting signal strength corresponding to that distance. For example, imagine a sensor in the middle of a room such as in the initial experiment in this thesis, with devices in and outside the room. In this case, assuming all the devices have roughly similar power levels, there should be a sudden drop in the signal strengths reported by the probe requests generated from outside the room when we look at their frequency distribution. Alternatively, if there was a stationary source of noise such as a phone shop next to our sensor where hundreds of phones regularly send probe requests, there should be a sharp rise in the of number of probe requests with reported signal strength corresponding to the distance between the sensor and the phone shop. Both of these changes can be identified by the 'breaks' in the distribution of the signal strength data, as demonstrated schematically in Figure 4.11. Identification of these breaks in the data should be carried out using traditional one-dimensional clustering algorithms such as 'jenks natural breaks', 'k-means', 'quantile' and 'hierarchical clustering', etc. which are usually used to find the class intervals in data. In simpler cases, the signal strength could be clustered into just 'high' and 'low' and the probes with low signal strengths could be ignored.

This approach has two primary advantages. Firstly, it does not rely on a predetermined threshold that has to be calculated with a representative sample, which is not usually possible in time-series data with such variability. Secondly, the methodology should apply for all the variations in micro-site conditions, since we are only looking for the relative breaks in the data and not for absolute values. For example, if the sensor is located inside an enclosure and all the signals are of generally lower

strength than usual, this method should still be able to find the distinction between the noise and the data from relatively near the sensors. The disadvantages of this method is that it might not work in situations where there are multiple sources of noise around the sensor, as they do not create a distinct pattern in their distribution.

Clustering with sequence numbers

There has been extensive research on extracting information about people from Wi-Fi probe requests in the past decade with feasible and favourable results. However, all of the methods used in the research depends on the Wi-Fi data having a primary unique identifier: a MAC address. When the MAC address is removed, or at least rendered non-unique, the established methods fail and cause significant risk to the infrastructure and commercial applications built around Wi-Fi data. As was shown in Chapter 2, various methods have been devised to overcome this anonymisation process including, but not limited to,

- *Profiling Manufacturers*: estimating the device model information from a known dataset of manufacturers and device behaviours [Martin et al., 2016]
- *Scrambler attack*: using another small part of the physical layer specification for Wi-Fi [Vo-Huu et al., 2016, Bloessl et al., 2015]
- *Timing attack*: where the packet sequence information along with information elements present in the probe request frame is used [Matte et al., 2016, Cheng and Wang, 2016].

A combination of these methodologies has been proven to produce de-anonymised globally unique device information[22]. Although these approaches are effective, sometimes even up to 90%, they usually result in a serious risk of breach of privacy of the users of the mobile devices by revealing their MAC addresses or by crossing ethical lines by tricking the devices into sending more information than they would ordinarily include in a probe request. Moreover, these risks are considered vulnerabilities by the computer security industry and are usually 'patched' in a reasonable amount of time, hence reducing their effectiveness in long term. As a consequence, it was necessary to explore methodologies for estimating the number of unique mobile devices from a set of anonymised probe requests, without the need to reveal their original device information.

Although the sequence number of the packet is not strictly unique to a particular mobile device, it was hypothesised that they can be used

[22] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, Frank Piessens, and Piessens Frank. Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pages 413–424. ACM, 2016. ISBN 1450342337. DOI: 10.1145/2897845.2897883; and Jeremy Martin, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik C Rye, and Dane Brown. A study of mac address randomization in mobile devices and when it fails. *arXiv preprint arXiv:1703.02874*, 2017

to estimate the number of unique devices. Vanhoef et al. [2016] used optional information present in the probe requests - Information Elements (IE) - along with the sequence numbers to successfully fingerprint the devices. This approach has become increasingly difficult as mobile phone manufacturers, especially Apple, have severely limited the number of information elements in the probe requests to curb such finger printing process. This problem affects established commercial solutions using Wi-Fi probe requests such as Blix, Walkbase, Euclid Analytics, and RetailNext etc. These companies solve the randomisation by combining Wi-Fi data with other sources of data such as cameras, lasers or infrared counters, but this is not possible for our research. More recently, another solution to the problem was proposed by Hong et al. [2018][23] whereby a *Hidden Markov Models* based trajectory inference algorithm was deployed. Unfortunately the research was limited to enclosed, exit-controlled public spaces such as shopping malls and railway stations, and therefore does not translate well to the open retail high streets studied in this thesis. As such, a novel method to suit the context of this research was devised.

[23] Hande Hong, Girisha Durrel De Silva, and Mun Choon Chan. Crowdprobe: Non-invasive crowd monitoring with Wi-Fi probe. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):115, 2018

The first approach taken was to establish a 'factor of randomisation': the ratio of the total number of randomised probe requests emitted to the number of unique MAC addresses in them. This factor was then used to adjust the counts when aggregating the randomised probe requests. As explained in section 3.5, the rate of probe request generation is highly variable and an approach which assumed a constant and stable rate of probe requests, was therefore not feasible. Moreover, since software and specification change frequently, it was surmised that this method was not feasible in the long-term. It was necessary to create a more general approach independent of the device model or manufacturer.

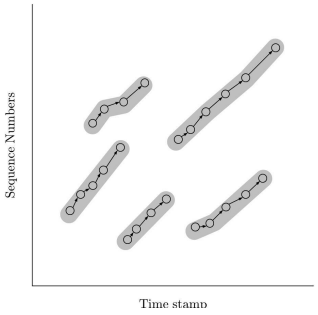


Figure 4.12: Thematic diagram showing the idea behind grouping sensors using their sequence numbers.

Resulting from the initial experiments explained in section 3.2, it was found that OUI and the sequence number of the probe request were the most promising information to achieve this. It was also observed that, when plotted against time stamps, sequence numbers show distinct streak patterns which could be isolated as single unique devices. Since only one probe request can be received at a time, it was possible to link them using a graph-based algorithm as illustrated in Figure 4.12. Such an algorithm would create a graph with the randomised probe requests whereby the nodes were the probe requests themselves. The edges were created between the nodes based on the following rules:

1. A link could go only forward in time.
2. A link could go from low to high sequence numbers.

3. A link could exist between nodes with a maximum time difference of α - time threshold.
4. A link could exist between nodes with a maximum sequence number difference of β - sequence threshold.
5. A node could have only one incoming link and one outgoing link, which is the shortest of all such possible links.

The first two rules arise from how the Wi-Fi data collection process works. The third and fourth rules create a kind of 2 dimensional moving window within which all the links are connected. The final rule simplifies the graph into strands of unique devices based on the assumption that the closest points within a window belong together. This, of course, will not work accurately when there are intersections between these strands. This solution could be made more accurate by calculating angular changes similar to techniques used in creating 'dual networks' in roads[24], but could disproportionately increase the amount of processing needed. Hence for this research, the former method which uses the shortest link was used.

[24] A Paolo Masucci, Kiril Stanilov, and Michael Batty. Exploring the evolution of london's street network in the information space: A dual approach. *Physical Review E*, 89(1):012805, 2014

After simplifying the graph conceptually, each connected component corresponds to a device generating probe requests periodically with increasing sequence numbers. A unique identification number was then assigned to the nodes based on the connected component of the graph they belonged to. This unique identifier was then used in place of MAC addresses for the aggregation of the anonymised probe requests. As discussed in section 3.1, the sequence numbers do not always increase as they get reset after 4096; thus, this method can lead to multiple unique identifiers being reported for a single device. This can be potentially solved by treating sequence numbers as a 'ratio' scale, while calculating distances between probe requests. Since a sample consisting of randomised probe requests sent by "Google" devices in the data collected from the initial experiments showed that only 0.5% of the sample had their sequence number reset in a given period, this effect has been deemed inconsequential and ignored in this research.

Calibrating with Ground Truth

Since proportion of mobile device ownership was an external uncertainty to this study and could arise from variety of spatio-temporal and demographic factors, the study aimed to solve the uncertainty by using a manual sample count at each location. An adjustment factor or an 'offset' was calculated for each location by comparing the sensor-based counts

and ground truth, similar to what was undertaken in the beginning of the project. This adjustment factor was then used to adjust the rest of the data reliably to reflect the ground truth in absolute numbers. On a project with a large scope, such as the Smart Street Sensor project, since this calibration applies in addition to the other methodologies, in addition to increasing accuracy of measurement in the short time, they can be carried out periodically at chosen locations to improve the quality of estimation over a long time.

The three methods – signal strength filtering, sequence number clustering, and manual calibration - together provide a complete methodology for converting the Wi-Fi probe requests into footfall. But the methods need empirical experiments for a successful implementation with real-world data. For example, the signal strength methodology we need to find the most suitable one dimensional clustering algorithm and for the sequencing method the values of threshold need to be calculated. These questions were answered by applying the methods on the data collected in the experiments and pilot study as detailed in the upcoming sections.

4.2.2 Oxford Street Experiment

The primary aim of the initial experiment conducted on Oxford Street, London was to collect data to validate the filtering and clustering methods against the scale and complexity of an open public area. It was also aimed at finding the algorithm which was best suited for the one dimensional classification of signal strengths as either 'low' or 'high', in order to filter out the background noise.

The first step was to create a base line count or 'raw count' without any cleaning procedures, whereby the probe requests were aggregated by their MAC addresses for every minute. This generated a continuous, minute-by-minute count of the number of people estimated to be near the sensor. It was assumed that each MAC address corresponded to a mobile device and hence a pedestrian. This preliminary 'footfall' count was then compared to the actual number of pedestrians recorded manually to check for robustness. The statistic - Mean Absolute Percentage Error (MAPE) – was used as a measure of robustness of the count, since it provided a simple and quick idea of how much the pair of time series data differed from each other. MAPE generally does not work with datasets with a significant number of data points which are not known, or those which contain zero values; however, because of the busy nature of the survey location, there were no such intervals without any footfall.

It was observed that the MAPE in these raw counts, when compared to the actual ground truth, was around 425%. This confirmed the presence of a large amount of noise in the data which may have been generated by the sources of uncertainties discussed in section 3.5. It also demonstrated the need for filtering the data before aggregating them into footfall.

Algorithm	Time (s)	MAPE
Quantile	0.002	27 %
K-Means	0.007	23 %
Hierarchical Clustering	172.520	9 %
Bagged Clustering	0.135	30 %
Fisher	3.034	30 %
Jenks Natural Break	556.279	30 %

Table 4.7: Comparison of clustering algorithms with a sample of 40000 probe requests

The probe requests were then classified as 'high signal strength' or 'low signal strength' using various one dimensional clustering algorithms. The algorithms used were as k-means, quantile, hierarchical clustering, bagged clustering, fisher and jenks natural breaks with the number of clusters set as 2. Due to the processor-intensive nature of some of these algorithms, only a sample of 40,000 probe requests were selected for this benchmarking exercise. For each exercise, the resulting probe requests were filtered only for those with high signal strength, and rest were discarded. As before, the filtered probe requests were then converted into footfall counts by aggregating them based on their MAC addresses, and subsequently compared to the manual footfall counts.

Two metrics were collected for each clustering algorithm: the time it took to classify all the data points in the sample, and the amount of MAPE in the resulting footfall estimates. The results are shown in Table 4.7. It was found that out of all algorithms, hierarchical clustering provided the least amount of errors. However, this and jenks natural break were designed to identify class intervals in much smaller datasets and were extremely resource intensive for practical use with a larger dataset. It was also found that the k-means algorithm gave the quickest result with the lowest MAPE, closely followed by the quantile algorithm. The cut-off point, or threshold, for the collected data with which we could classify the probe requests as high and low was found to be -71 dBm, using the k-means algorithm. When the data were aggregated after this filtration process to remove all the probes with a 'low' signal strength, it resulted in a footfall count with a MAPE of 30%. This was extremely encouraging considering the magnitude of improvement. However, it was not certain if this filtering process was removing noise only from outside, or if it had any kind of independence from the configuration of the sensor at the location. These concerns needed to be addressed with a

larger survey with multiple locations, as discussed in the pilot study.

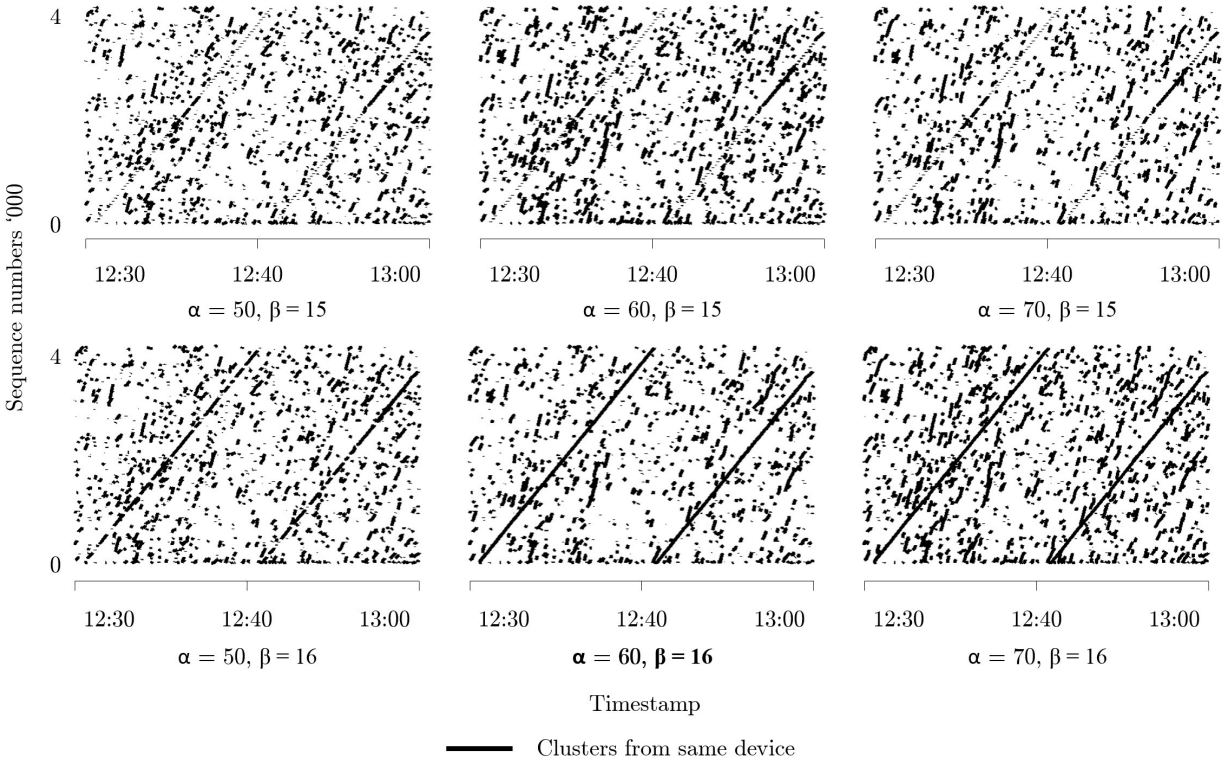


Figure 4.13: Finding the optimum time threshold (s) α and sequence threshold β through trial and error.

With the filtering method validated, the next step was to identify the probe requests which were generated by the same device irrespective of the MAC randomisation using their sequence numbers. The graph theory based algorithm defined earlier was employed and each local probe request was assigned an alternative unique identifier or signature independent of their the MAC addresses. Since a baseline for the nature or frequency of the MAC address randomisation process could not be established, the surveyor's mobile device was used as a reference. As the surveyor's device was being actively used to count pedestrians with its Wi-Fi module kept active without establishing connection to any network, it was known that the device was continuously probing for new networks. Moreover, since the screen of the device was switched on with constant taps the frequency of these probe requests were higher than normal. It was also known that the OUI of the device corresponded to the vendor - 'Google' – and that the device was regularly randomising its MAC address. All in, it served as an excellent reference point with which it was possible to determine if the clustering method has worked.

The algorithm required two parameters which needed to be determined empirically - sequence threshold and time threshold - which was

done using trial and error as described below. The clustering process was done repeatedly with increasing values for both thresholds in increments of 1, and for each increment, the resulting clusters were examined to see if the data from the reference device were clustered into one device. The minimum possible time and sequence thresholds at which the algorithm clustered the reference device properly without over clustering the other probe requests, or under clustering as multiple devices, is illustrated in Figure 4.13. It can be observed that the threshold for time α and the threshold for sequence numbers β, are 16 seconds and 60 respectively.

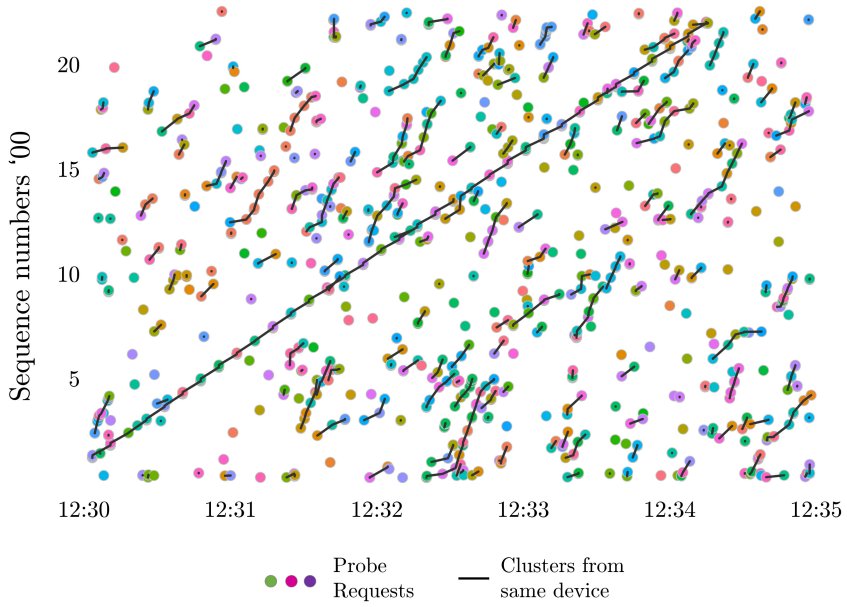


Figure 4.14: Sample showing the result of sequence numbers based clustering algorithm on data collected at Oxford Circus, London.

Figure 4.14 shows the results of this clustering process on a small set of randomised probe requests collected in this experiment. The probe requests with different randomised MAC address are shown by the coloured dots and the lines joining them show that those probe requests were clustered together by the algorithm and are most likely generated by the same device. The data were finally aggregated as before, but with this device's signature rather than the local MAC addresses. This resulted in a footfall count with a MAPE of -18% compared to the manual count. It is important to notice that this clustering was undertaken on top of the signal strength filtering, and only for the probe requests with randomised MAC addresses. A comparison of minute by minute counts resulting from different filtering processes along with the ground truth is shown in Figure 4.15, and illustrates the promising effectiveness of the methods.

To summarise, the data from the initial experiments suggest that both filtering using signal strength and the clustering using sequence numbers worked well on complex, real world data and resulted in fairly accurate pedestrian counts with a MAPE of 20%. It was also found that 'k-means' and 'quantile' are the best algorithms for clustering signal strengths, and the optimum thresholds for time and sequence numbers for the clustering algorithm were around 16 and 60 respectively.

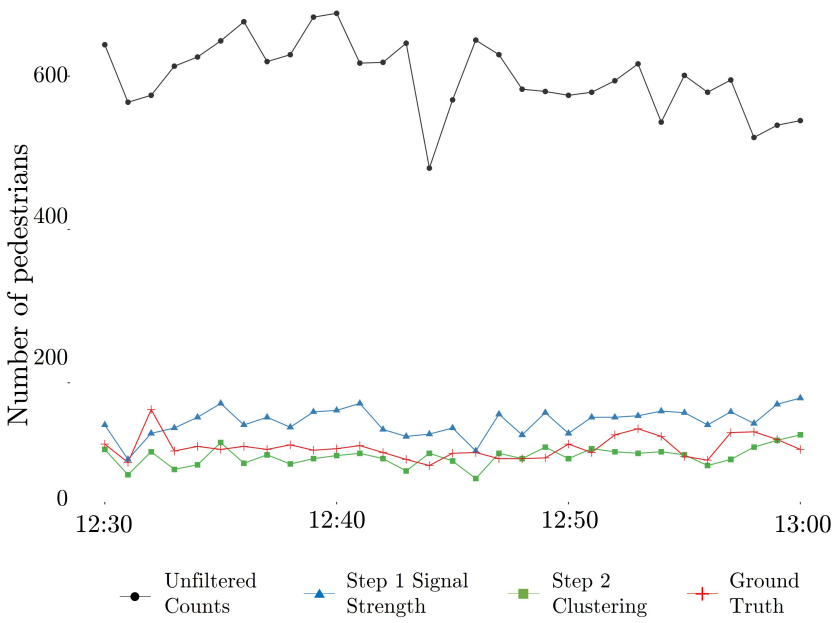


Figure 4.15: A comparison of estimated footfall at Oxford Circus during various stages of filtering with the actual manual counts.

4.2.3 *Pilot Study*

With encouraging results from the initial experiment, the next step was to check if the methods worked on the various locations where sensors were installed in different configurations using the data from the pilot study. The distribution of the signal strengths of all the probe requests collected at each location were created and were compared to the corresponding sensor configurations at these locations. When visualised as density plots as shown in Figure 4.16, they show a clear relationship between the distribution of the signal strength and the distance and complexity of the source of noise at each location. It was observed that when there is clear distinction between the source of noise where stationary devices were generating randomised local probe requests, the signal strength distribution shows a clear distinction between high and low values. For example, at location 5 where the noise is generated by a phone shop next door, a significant spike in the number of probe requests generated

with similar low signal strength was also reported. Whereas, location 2 - a restaurant located next to a public square with seating on both sides of the sensor - shows no such discernible distinction of any sort. This demonstrates that filtering using signal strength works, but at the same time depends heavily on the assumption that the sensor is installed in such a way that the field of measurement is clearly 'visible' in terms of distance from it. Figure 4.16 confirmed the intuition that data collected at location 2 and 4 would be harder to clean than the ones collected at locations 1, 3 and 5. The signal strength threshold, calculated using k-means algorithm, for all the locations except for 2 were between -72 dBm and -70 dBm - very much in line with the findings of the initial experiment. This also introduced the possibility that -70dBm could be used as a rule of thumb for filtering noise at a general location unless it faced specific challenges like location 2. It is important to note that as the aim was to compare the counts with manual counts, the data used to calculate the threshold pertains only to the time when manual counting was undertaken at these locations. Like before, the data were aggregated using MAC addresses after removing points with low signal strength and compared with manual counts. The results are shown in Table 4.8. In this exercise probe requests with MAC addresses which repeated within a 15 minute window were also removed.

We observed that the MAPE at locations 1,4 and 5 was reduced to -19% to 150% from the original 250% to 500%, making this an ideal candidate for a quick and easy cleaning procedure for most practical applications. Locations 2 and 3 were found to be particularly tricky: the former had the propensity to overestimate the footfall, and the latter underestimated it. This could also be attributed to the configuration of the sensor at the particular location. Location 3 was particularly interesting as it is the only location with no source of stationary noise and almost all the probe requests collected at this location should be coming from genuine footfall. It was observed that the filtering needs to be less aggressive in locations without any obvious source of noise to prevent underestimating footfall at these locations.

The next step was to test the sequence number based clustering algorithm on these locations. The sequence number clustering algorithm was run on the local MAC addresses to cluster the probe requests; the resulting device signatures were used to aggregate them into footfall counts. The results showed that this process further reduced the MAPE to almost 13% - 300% on all the sensors with a clean configuration. Locations

3 and 2 were still outliers due to their complex configurations. The final step was to test the simple manual calibration process: an adjustment factor was calculated for each location as the ratio between footfall estimated from the sensor after processing and the actual pedestrian counts. This adjustment factor was used to adjust the counts. It is important to note that the adjustment factor and MAPE were calculated using an interval from a different location than the manual counts. This process further reduces the MAPE to 10% - 50% while taking care of the over-counting problem in Location 3. Location 2 is still not as accurate as the MAPE after all the processing that was done, which highlights the limitations of the methods discussed. Figure 4.17 shows the effectiveness of these cleaning procedures at each location.

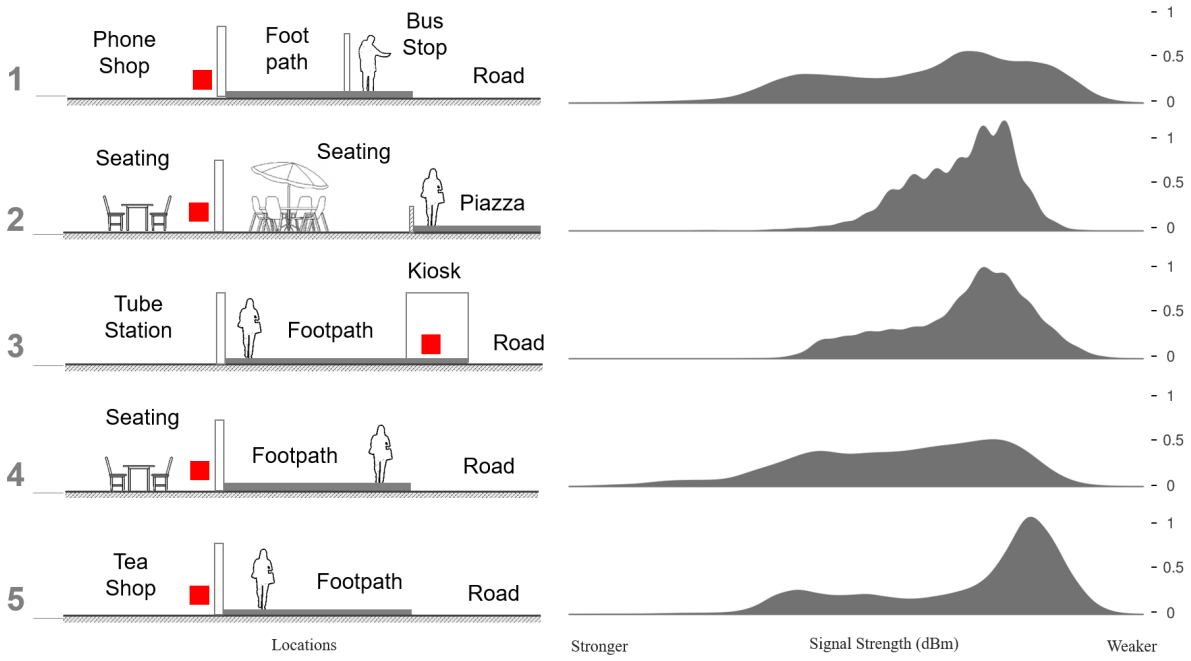


Figure 4.16: Distribution of signal strengths at locations covered under the pilot studies along with the corresponding configurations of the sensors.

To summarise, the pilot study confirmed the findings from the initial experiment by showing that the signal strength based filtering is effective and provides a quick and easy way to clean out the noise, when used along with the sequence numbers based finger-printing technique. It also demonstrated that the sensors with no discernible stationary source of noise tend to under-count pedestrians, and therefore require that calibration is done using manually collected data. In contrast, sensors with seating next to them significantly over-count footfall. However, the study also proved that through the process of cleaning, the counting errors can be reduced substantially resulting in the sensor based counts

being accurate within 10% of the ground truth.

Sensor	Signal threshold (-dBm)	Adjustment factor	MAPE before filtering (%)	MAPE after filtering (%)	MAPE after clustering (%)	MAPE final adjusted count (%)
1	-70	1.25	259	22	-13	9
2	-74	0.51	928	396	206	55
3	-72	1.60	87	-19	-31	10
4	-70	0.88	498	142	52	33
5	-72	0.80	473	84	38	11

Table 4.8: Results of footfall estimation at each location as Mean Absolute Percentage Error (MAPE) after each step of the filtering process.

4.2.4 *Smart Street Sensor Project*

In addition to solving problems arising due to privacy oriented decisions and figuring out methods to enhance Wi-Fi based analysis, one of the primary objectives of the research was to solve the problem faced by the Smart Street Sensor (SSS) project in respect of the explosion of MAC randomisation. The number of probe requests and unique MAC addresses nearly tripled in one week in the autumn of 2017, which essentially made the data unusable and resulted in an extreme risk to the project's feasibility. The methodologies discussed above were perfect for the SSS project, and when implemented it would improve the project's long term feasibility immensely. However, being designed from a commercial point of view by the data partner, the SSS project's implementation posed significant challenges in adapting to the methods as well.

- *Lack of data* - The data collected by the SSS project is optimised for transferring large amounts of data with the least possible cost. Hence, the project collects only the most important fields. For example, sequence number, information elements, SSID, etc. are not available in the dataset.

- *Aggregation* - To further optimise the size of the data transfer, the data points are aggregated at the device level every five minutes. Hence the time information has a resolution of 5 minutes and the signal strength of packets with same MAC addresses were summarised to either the maximum or minimum values observed in that interval.

- *Cost* - When changing the design of the project, the major challenge faced was the cost associated with every change. For example, due to the size of the project, changing the amount of data transferred even marginally translated into a significant increase in cost in terms of internet bandwidth, or the introduction of a large change in the

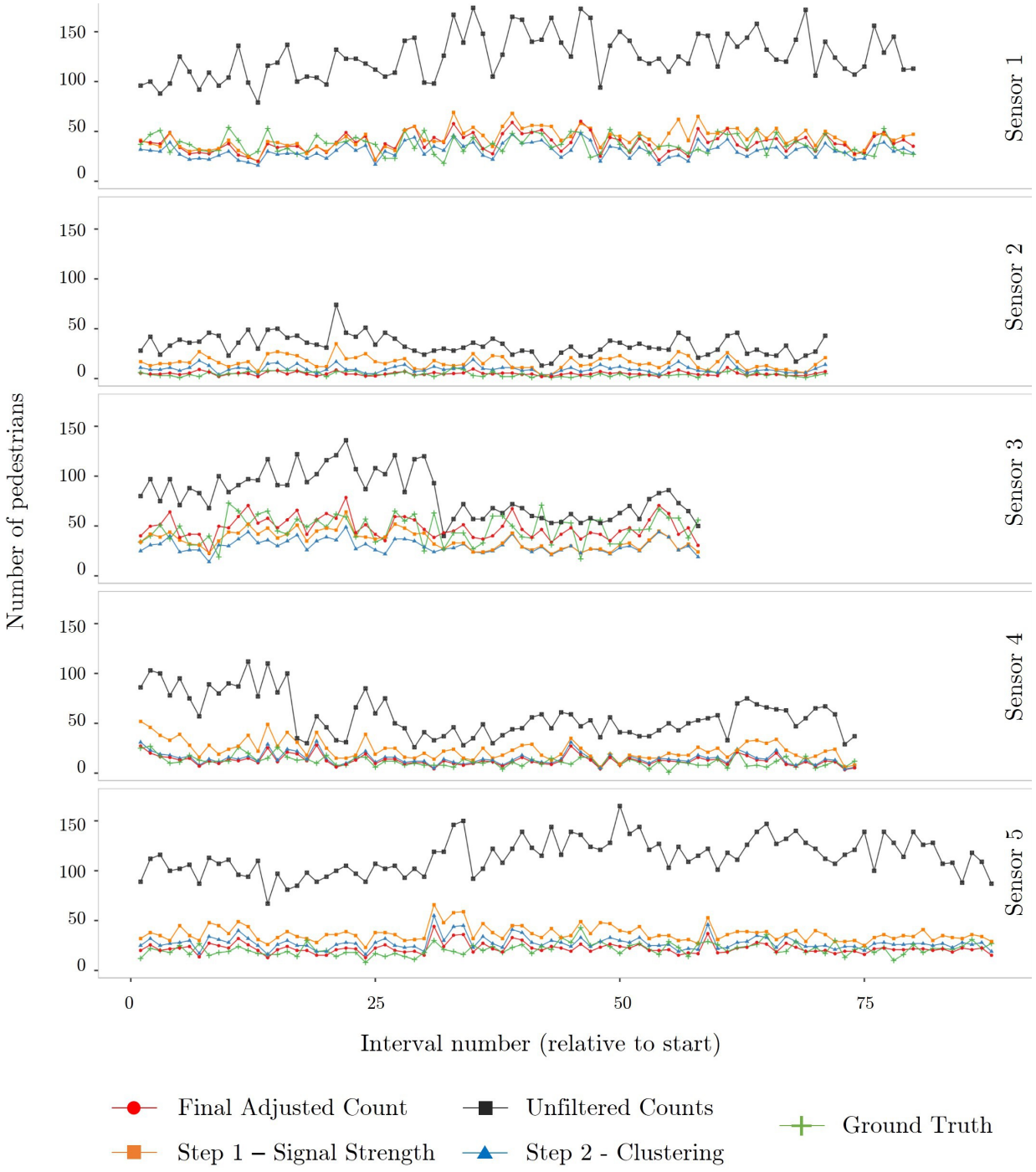


Figure 4.17: A comparison of estimated footfall at pilot study locations during various stages of filtering with the actual manual counts.

project also involves a cost increase in terms of development, testing, deployment and testing.

The above challenges made the sequence number based clustering impossible to use with the data available through SSS sensors; as such, only the signal strength based filtering could be applied. The signal strengths did not have the same granular information in them as the pilot study which may have greatly affected the output of the k-means algorithm. Nevertheless, an attempt was made to use the signal strength based filtering on the data generated from the SSS sensors. The data collected at the location where the pilot studies were conducted, along with manual counting, were extracted out of the SSS project and the filtering methodology was applied to the data. Since we only have the minimum signal strength for all the probe requests that were compressed, the data were weighted using the total number of probes that constituted the aggregated packet. Figure 4.18 shows that the difference between the signal strength distribution measured by the sensors used in the pilot study and the corresponding sensors from Smart Street Sensor project. It can be observed that the pilot study sensors collected data which shows a sharp change in the number of signal strengths after -70dBm comprising mostly of local MAC addresses, while the smart street sensors show a much more even distribution in both local and global MAC addresses. Locations 4 and 3 are excellent examples indicating this difference in distribution, where the signal strengths of randomised probes were more normally distributed in the Smart Street Sensors than in their counterparts. This was expected to pose a significant challenge to our filtering methodology, especially in locations with configurations similar to 4 and 3.

There was an expectation that the change in the distribution might cause vast differences in the results from these two exercises. However, when the probe requests collected by Smart Street Sensors were subject to filtering using the k-means algorithm and aggregated based on the MAC addresses after first removing the requests with low signal strength, the results were surprisingly close to what was reported by the pilot studies. Figure 4.19 shows the results of the attempt to clean the smart street sensor data using the signal strength filtering technique, and compares them with the results of the pilot study. It is important to note that the sequence number based clustering has not been carried out in either of the datasets as the sequence number is not available in the Smart Street Sensor data. In general, it was observed that the results from the

pilot study and the Smart Street Sensor project were close compared to the manual counts after the signal strength based filtering. At Location 1, the pilot study sensor was massively over-counting on February 12 which could be attributed to a bug in the hardware: the additional Wi-Fi module which was installed for troubleshooting was generating probe requests in large amounts in response to very strong signal strengths. This was fixed later and the next manual count performed on February 21 confirmed that the sensor was fixed. At Location 2, the pilot study sensor resembled reality much more than the Smart Street Sensors, but footfall at this location was very low to begin with and both series showed similar trends and could be adjusted with a simple factor. It was also observed that, similar to the results of the pilot study, Location 3 led to under-counting, Location 4 led to over-counting, but Location 5 showed the best results owing to the clean configuration of the sensor in relation to its surroundings. Although the results are similar, they are still far from ground truth counts due to the MAC randomisation process and are also vulnerable to long-term changes, as shown in Figure 4.10. This necessitates an alternative to the sequence number clustering for Smart Street Sensor data.

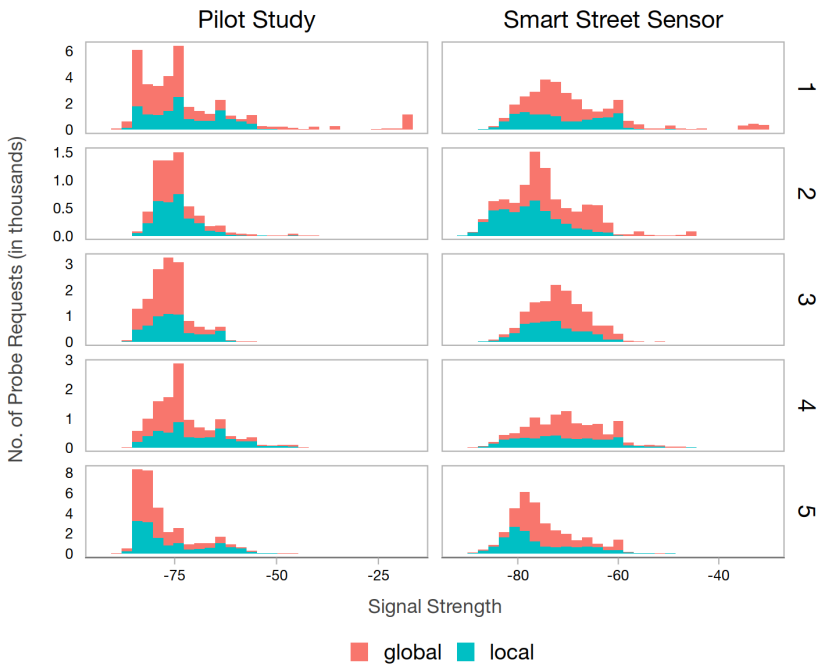


Figure 4.18: Comparison between distribution of signal strengths in probe requests collected by

Device to Probes Ratio

While searching for an alternative to the sequence number algorithm

with the limited data offered by the Smart street sensor, a much simpler approach was discovered. This approach entailed found that the number of probe requests generated by devices which do not randomise their MAC address can be used to estimate the devices which do randomise them for a specific interval. Assuming that, on average, all devices emit a certain number of probe requests in a given time interval, we can estimate the number of devices that are randomising as below,

$$Randomising\ devices = \frac{Non\ randomising\ devices}{Non\ randomised\ probes} \times Randomised\ probes$$

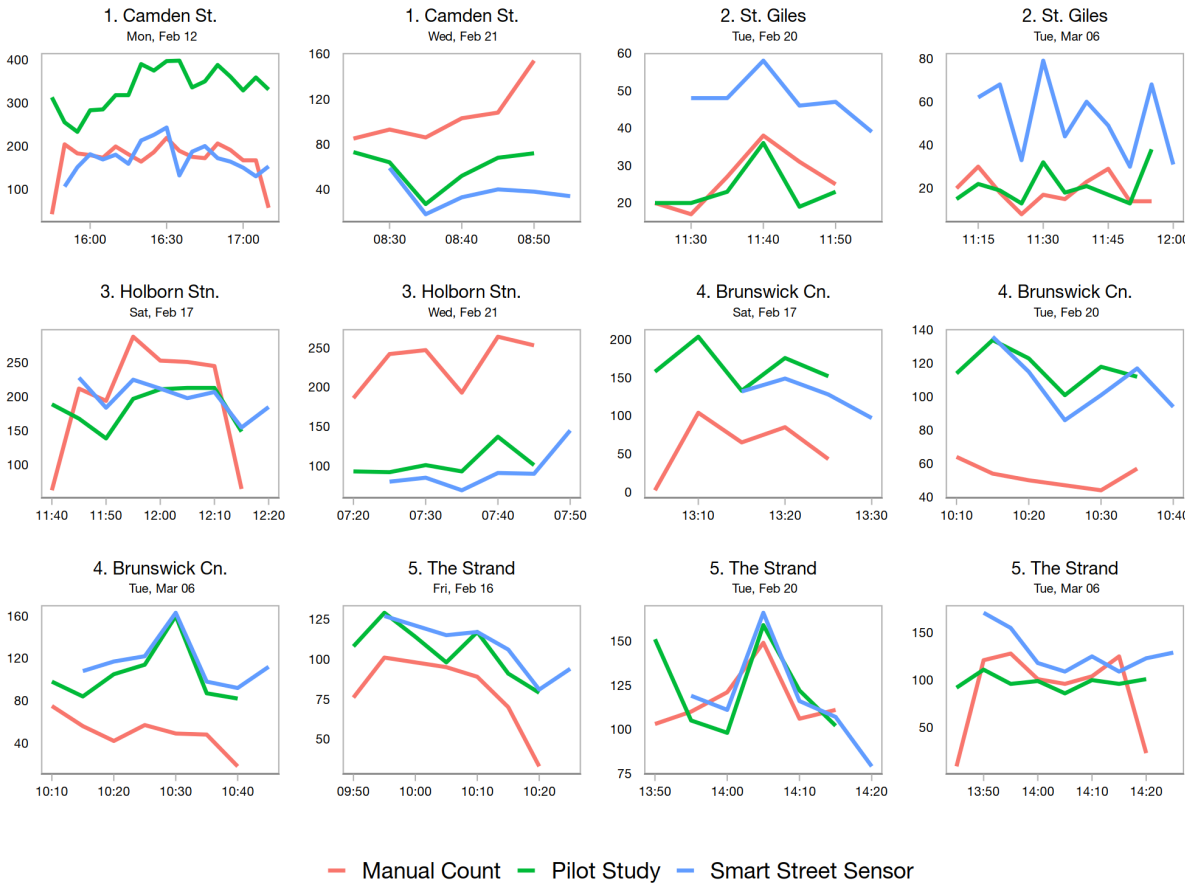


Figure 4.19: Comparison between the footfall estimates from Smart street sensor and pilot study after filtering probe requests of low signal strength along with manual footfall counts.

Although previous work on the probe request frequencies of different mobile devices demands scepticism over how well this method may work, the results were found to be encouraging. In real world data, the rate of generation of probe requests between various mobile devices was not constant. By encapsulating the problem in a small time interval - 5 minutes - and observing the corresponding non-randomising mobile phones, the behaviour of the randomising phones could be confidently

estimated. Moreover, if the randomisation techniques change in the future, the non-randomising phones should also alter, thus making such an approach robust against any future changes. Figure 4.20 shows the methodology applied for data from locations across Cardiff. The chart shows average weekly footfall estimated at each location before and after adjusting the number of randomised MAC addresses using the 'compression' ratio described above.

It was observed that, when adjusted, the resulting long term trend not only avoided the huge inflation experienced by the unadjusted estimate, but preserved the relative changes and trends and even showed seasonal variations. The only expected disadvantage of this method is the adoption of randomisation techniques by manufacturers over the time. Since the method depends on the number of devices that do not randomise their MAC address, when they decrease significantly in number, the method will fail. But for the time being and for the data which have been collected over the past 3.5 years, when combined with filtering using signal strength and other adjustment methods described in section 4.3, this method solves the uncertainties sufficiently. Moreover, the method also potentially provides researchers with a reliable and sufficiently accurate estimation of footfall from the data collected by the Smart Street Sensors.

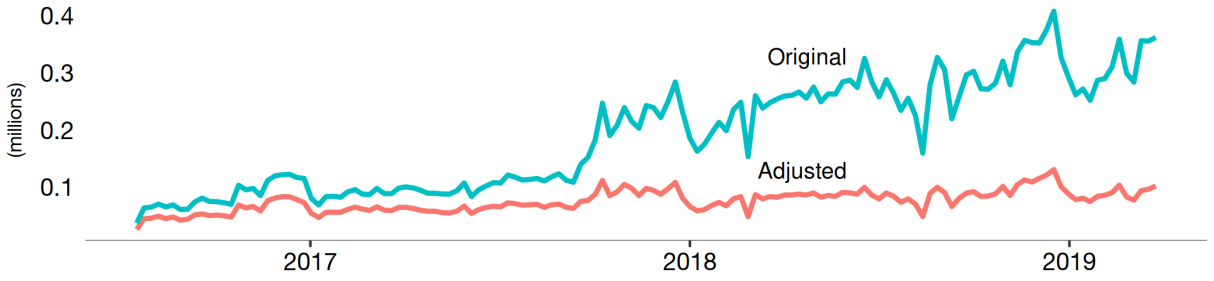


Figure 4.20: The result of the adjustment using device to probes ratio in non-randomising devices shown through average weekly footfall estimates for locations in Cardiff.

4.2.5 *Conclusions*

In this section, the uncertainties in the data collected through Wi-Fi - the ambiguity of the field of measurement and anonymisation of devices using MAC address randomisation - were discussed, alongside the extent to which these uncertainties affect the datasets. Two techniques were proposed to tackle these sources of uncertainty. The first technique involved using the strength of the signal reported in the probe requests

to cluster the requests into 'high' and 'low' with the help of one dimensional clustering algorithms. The second technique used the sequence numbers of the probe requests to group together probes generated by the same mobile devices using a graph based clustering algorithm. The effectiveness of both of these techniques were tested against the data which were gathered in Chapter 3.

The data collected from the initial experiment on Oxford Street, London showed that filtering the probe requests using the signal strengths works very well. In this case it reduced the MAPE from 435% to 20%. This case study showed that k-means is the most suitable method of doing the clustering, while the threshold for this specific dataset is -70dBm similar to the others collected through initial experiments. The case study also showed that fingerprinting unique mobile devices using sequence numbers is possible even in real world scenarios. Through trial and error and a reference device, the ideal values for the parameters for the algorithm - time threshold(α) and sequence threshold(β) are 60 and 16 respectively.

With the help of the pilot studies, it was also shown that signal strength based filtering does not always work efficiently everywhere. The distribution of signal strengths were found to vary widely when the sensors where installed in different configurations. The counts were found to suffer from a lot of over-counting when there are multiple sources of noise located close to the sensors, and they also under-count considerably when used in a location where there are no significant sources of noise. This emphasised the need for evaluating the site conditions and deploying manual calibration when utilising these techniques in real data. The sequence number based fingerprinting was also demonstrated to work in all these locations. This, along with signal strength and manual counts, reduced the error in the estimation from 470% to 10% in a 'clean' location with proper sensor configuration.

Unfortunately, the data from the Smart Street Sensor project was found to be insufficient to be able to effectively utilise these techniques, as the data were focused, aggregated, and prohibitively expensive to change. Surprisingly, the signal strength data which were aggregated for five minutes for each MAC address by the minimum value, worked as well the detailed per packet data collected in the pilot study. Although this, along with manual counting, can help in reducing the error, it does not solve the problems introduced by MAC randomisation, especially the long-term changes. An alternate method using the ratio of probes to

non-randomising devices was proposed and was found to be extremely efficient in eliminating the noise created by the randomisation. The sample data from Cardiff showed that this method not only removes the massive increase in footfall estimates experienced since September 2017, but it also preserves the seasonal variations in the data thus enabling the data to be used for research into long-term phenomena.

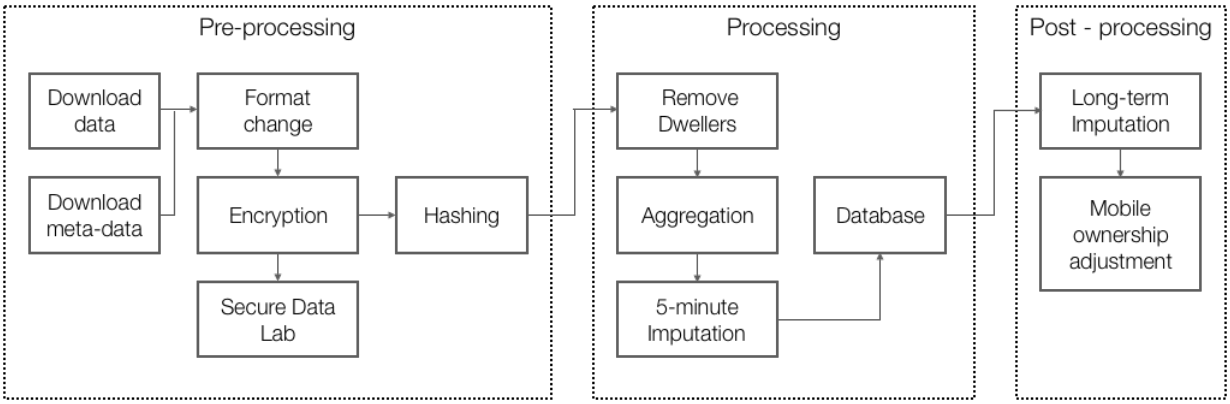


Figure 4.21: The complete data processing pipeline which takes in raw probe requests from Smart Street Sensor project and outputs footfall estimations.

4.3 *Data pipeline*

After the discussion of the toolkit and the methods, the final step of this research was to convert the Wi-Fi probe requests to footfall. This meant devising a processing pipeline which combines the 'data toolkit' and the methods together, takes the probe requests generated by the Smart Street Sensor as input and generates a best possible estimation of footfall estimations as output. Figure 4.21 illustrates the pipeline that was devised as part of this research. The code which implements this pipeline is detailed in Appendix 7.3. The pipeline comprises of three parts,

- *Pre-processing* - where the data gets transformed and modified for security and convenience.
- *Processing* - where the data gets converted to estimated ambient population.
- *Post-processing* where the general population estimate gets adjusted to better reflect information that was sought - footfall in this case.

4.3.1 *Pre-processing*

The aim of *pre-processing* was primarily to convert the dataset comprising of probe requests into a form which can be processed quickly and

conveniently. In the case of the Smart Street Sensor project, the data is retrieved from the *Azure* store and converted from JSON format into CSV file for each location on a given day. These files were stored in directories within a file system hierarchy of year, month and date providing us an efficient storage system as discussed in Section 4.1.

The secondary aim of pre-processing was to anonymise the data using cryptographic hashing and encryption to protect the privacy of the users. The MAC address field in the dataset was converted into an unique hash value using SHA256 algorithm to avoid linking this dataset with other sources of MAC address data and a random salt value was introduced every week to avoid timing attacks where sufficiently long term information on hashes could be de-anonymised by correlating the precise time they occur in other datasets. Although this is not theoretically fool-proof, this made sure that dataset could not used to personally identify the users within practical means. Finally, the non-hashed raw data were encrypted using RSA algorithm and physically transferred to an isolated secure facility for storage. Thus preserving the unmodified information for further investigation, if required.

4.3.2 *Processing*

The processing of the data involved the implementation of the methods discussed in Section 4.2. First the probe requests were separated into global and local based on the OUI present in them. The non-randomised, global probe requests were then aggregated into ambient population estimation employing the following steps,

1. Signal strength filtering - as discussed in section 4.2 the probe requests with signal strengths less than the *threshold* (Calculated dynamically from the previous 24 hour period) were filtered out.t
2. Removal of dwellers - The probes with MAC addresses which are repeated within the past 30 minutes were removed to eliminate noise caused by devices which are dwelling around the sensor for a long time.
3. Aggregation - The remaining probe requests were aggregated by their MAC addresses to arrive at the count of unique MAC addresses at every five minute interval.

The above process results in the number of devices that were present around the sensor which did not randomise their MAC addresses. The

ratio between this 'count' and the original number of probes requests with global MAC addresses gives us the *compression factor* as discussed in Section 4.2. This factor was calculated for every interval and in turn used to calculate the number of devices which randomise their MAC addresses. The sum of the both counts gives us the estimate of ambient population at the locations for five minute intervals. As the final step, the gaps in the data which are shorter than fifteen minutes are filled in by imputation based on *Kalman smoothing* resulting a more continuous dataset.

4.3.3 *Post-processing*

The post-processing involved adjusting the estimated ambient population further to match it to the real footfall counts as closely as possible. The primary steps involved were, adjusting using manual counts, imputing missing data and adjusting for increase in mobile device ownership. The manual adjustment was done through a global adjustment factor which is calculated for each location by comparing the footfall measured with the sensor to the footfall counted at the corresponding locations manually. The manual counting were done for short intervals ranging from 15-30 minutes immediately after the installation of the sensors and are refreshed for selected locations yearly. This adjustment factor, which could be greater or less than 1, was then applied on sensor counts to account for over- and under-counting at these locations. This adjustment removes the noise caused by the location specific configuration. In addition to this the counts since the beginning of the project were offset by 0.2% per week to account for the increasing mobile phone ownership over the years in the United Kingdom[25]. Finally, long term gaps in term of hours, days and weeks were filled using a seasonally decomposed method hierarchically where the daily and weekly periodicity of the counts were taken into account during the imputation process. This ensured the final counts produced from the processing are continuous and standard enough for carrying out further research using them.

[25] Deloitte. Mobile consumer survey - united kingdom, 2018. URL https://www.deloitte.co.uk/mobileuk/

5
Applications and Visualisations

This chapter serves as a gallery of possible applications that can arise from the data generated from the Smart Street Sensor project. As was seen in Chapter 2, availability of granular, longitudinal data on the movement and distribution of people at such spatial extent has numerous uses in various fields of study. This chapter first starts by looking at the use of the data in understanding the broad footfall landscape of the United Kingdom (UK), deriving sample insights on how retail footfall in the UK has been performing for the past couple of years along with some sample analysis from the national level to individual locations to understand the nature and change of footfall. It then demonstrates couple of ways of detecting events and their effect on places from the change in the footfall volumes. Finally the chapter briefly describes a way to calculate the flow of pedestrians between the locations just from the highly granular footfall volume using a probabilistic approach.

5.1 Footfall Indices

5.1.1 UK Footfall Index

One of the broad questions that arises when such footfall data is available is about the general national trend of footfall on retail high streets. This national 'footfall index' is not only important for the retail industry but also for various other purposes such as policy making, economic forecasting, etc. Since the footfall has a weekly periodicity, a standardised index could be arrived from sensor based data by aggregate them at a national level and find the average footfall at every location at every week. Although this is a simplistic measure, it does a good job in describing the changes in the retail footfall in the UK as a whole. From the last two years, it can be observed that retail footfall in UK started at its lowest in the beginning of the year and increased steadily until spring. The high

footfall lasted through the summer months before going down steadily towards the end of the year.. This trend changed around the fall months and the footfall reached the highest in the first two weeks of December and fell back to the lowest in the last two weeks. This form the yearly pattern of retail footfall in UK is illustrated in Figure 5.1 which shows the weekly footfall index of the UK from 2017 to 2018.

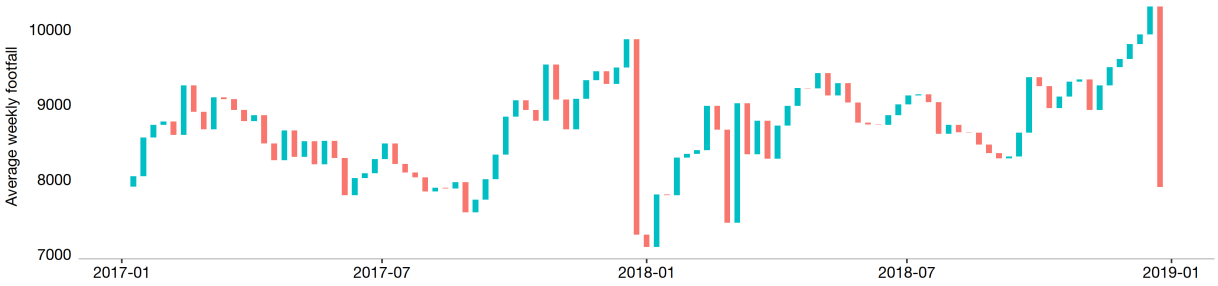


Figure 5.1: A weekly footfall index for United Kingdom showing the change in footfall from 2017 to 18

In addition to showing larger trends, this footfall index also showed sudden short term changes. One such as example was the storm in February 2018 named - "Beast from the East", which resulted in a week of unusually low footfall experience across the UK. This 'footfall index' can serve as a measure to give the overall outlook for the retail activity in the country at any given week.

5.1.2 *City-wise footfall index*

Being derived from location-wise data, the index could be calculated for geographic extents as well. For example, Figure 5.3 shows the spatial distribution of footfall change in the UK towns between 2017 and 2018. On a glance, it can be observed that the small southern towns such Ipswich, Staines, Southend by Sea and Plymouth had a good amount of growth in footfall while the towns such as West Bromwich, Derby and Warrington have a decline in footfall. In addition to long term changes, from the footfall data, even insights on short term changes could be derived. For example, Figure 5.4 shows the change in footfall from the previous year for towns across UK for just the months of April and May 2019. It can be observed that April 2019 has been slower than last year in most towns across UK while May 2019 has been actually better than last year. This kind of granular insights into trends in footfall could be valuable for local authorities who can measure and monitor the health of their retail areas closely. The difference in even smaller intra-day patterns in cities could be derived from footfall data which could show the nature of their economies Figure 5.5 shows an average daily footfall pattern for

9 cities in the UK.

5.1.3 *Location profiles*

One of the most interesting insights that can be derived from the dataset is the detailed knowledge of the locations themselves. The data can not only reveal how popular a location is but also the exact times when the location is popular. The patterns in the usage of the locations could also reveal the function of the place giving us the opportunity to measure their change through time as well. For example, Figure 5.2 shows the daily footfall profile of three locations in London for two weeks in 2019. It can be observed that all three locations have completely different patterns of usage. Leicester Square was mostly a evening destination where the footfall peaks around evening while Regent street is a mostly office location with three distinct peaks corresponding to morning commute, evening commute and lunch. These insights can be crucial for retailers operating in these places for optimising their business operation in terms of store opening times, scheduling shifts etc.

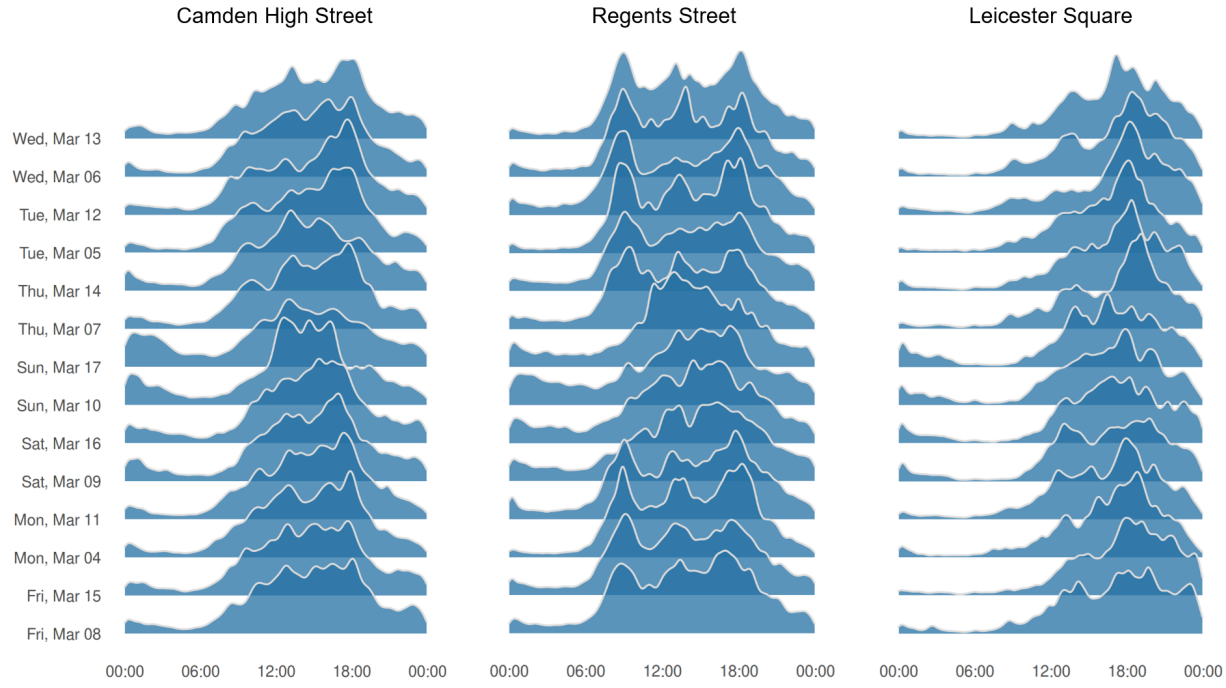


Figure 5.2: The profiles can be tracked longitudinally to reveal nature and change.

Another way to understand the evolution of a place over time is to look at the patterns of its usage over the corresponding period. For example, Figure 5.6 shows a 'footfall calendar' for Old street, London which traces and visualises the evolution of the place for the first half of 2018.

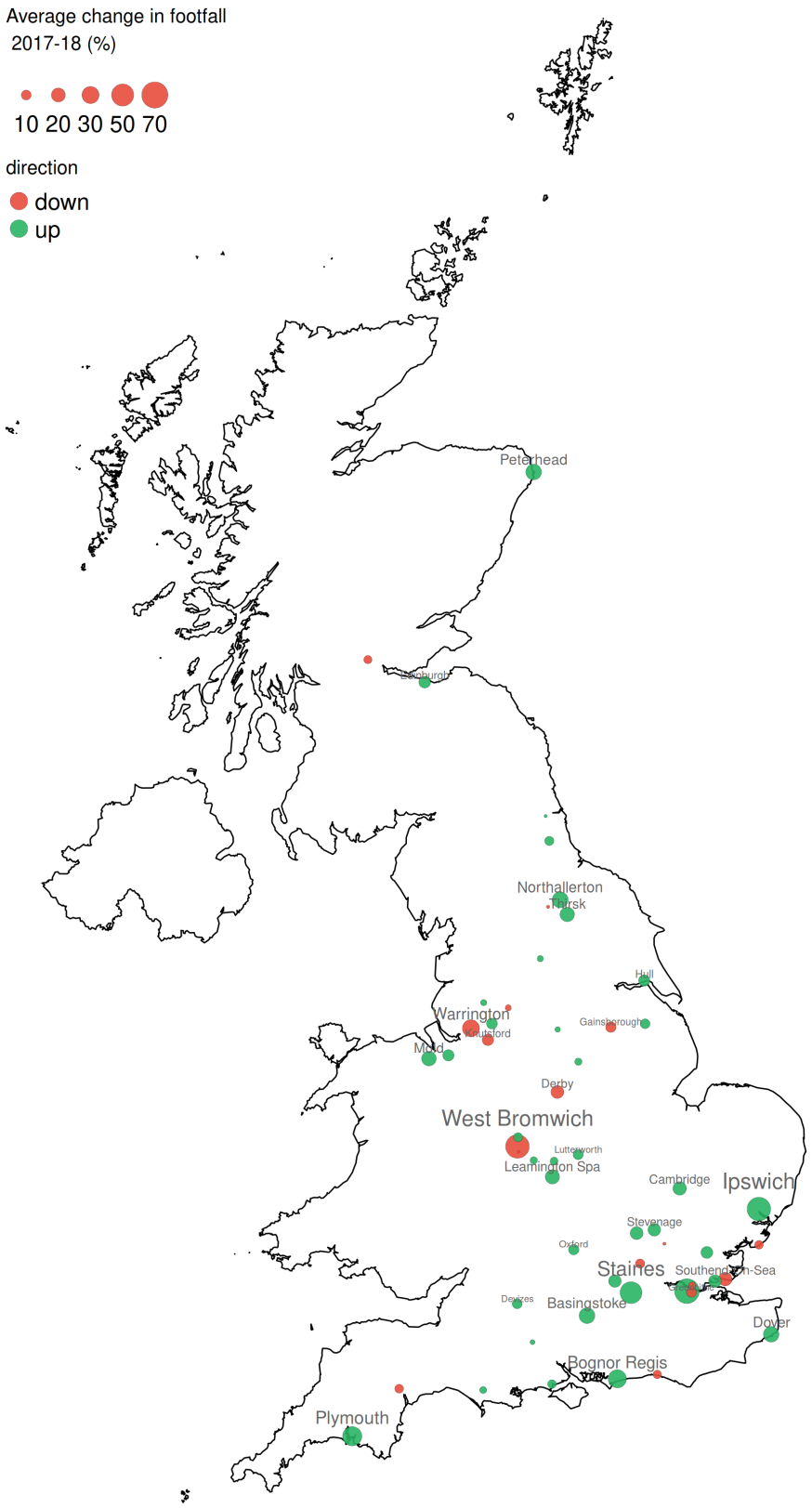


Figure 5.3: The change (%) in average weekly footfall of towns across the UK in 2018 compared to 2017.

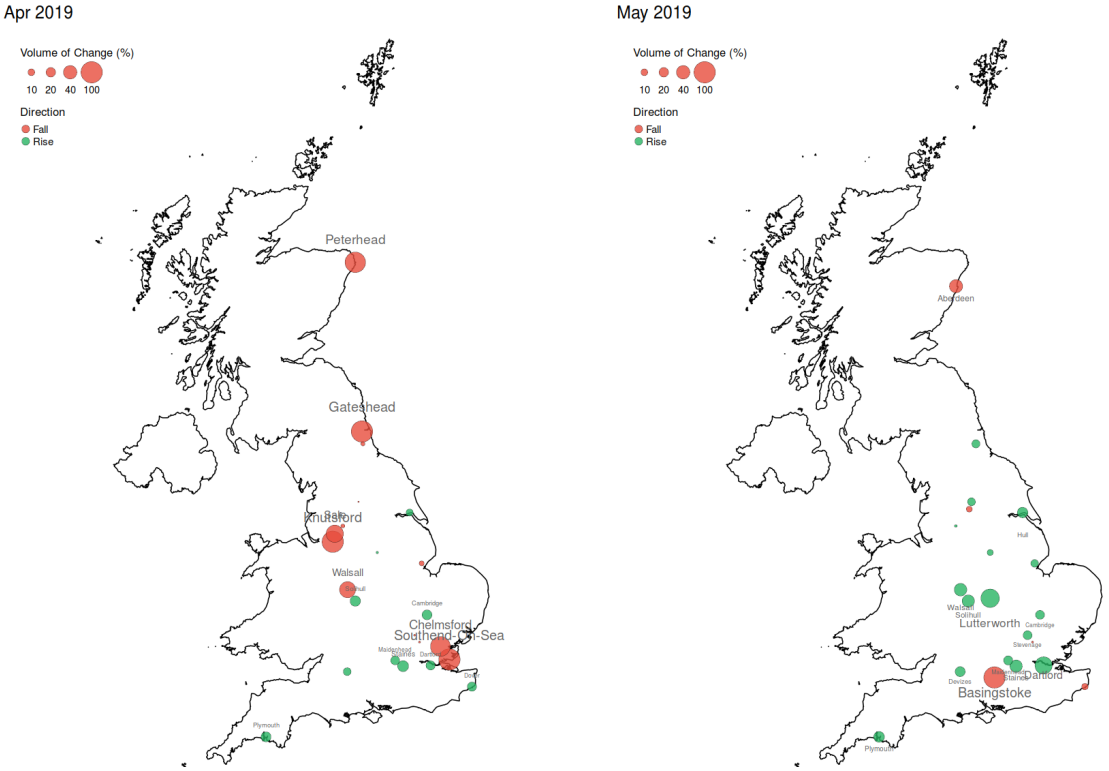


Figure 5.4: The change (%) in monthly average footfall in towns across the UK in April and May 2009.

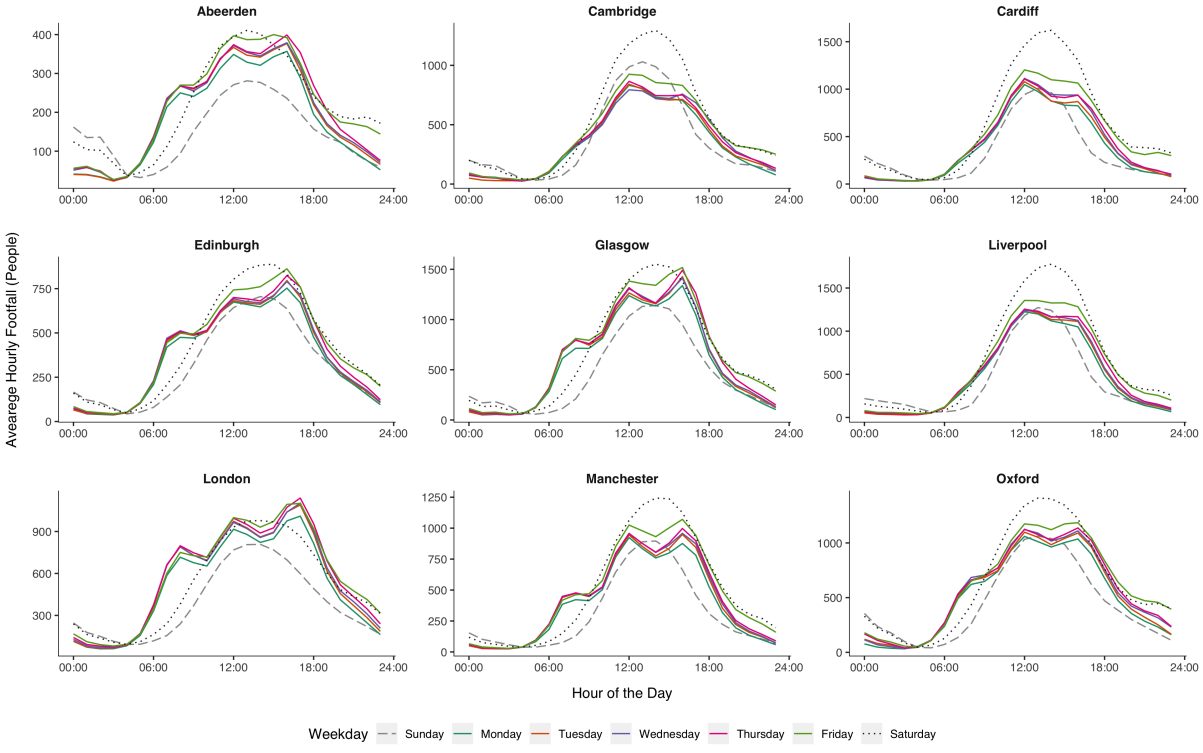


Figure 5.5: Intra-day footfall profile of major cities in United Kingdom

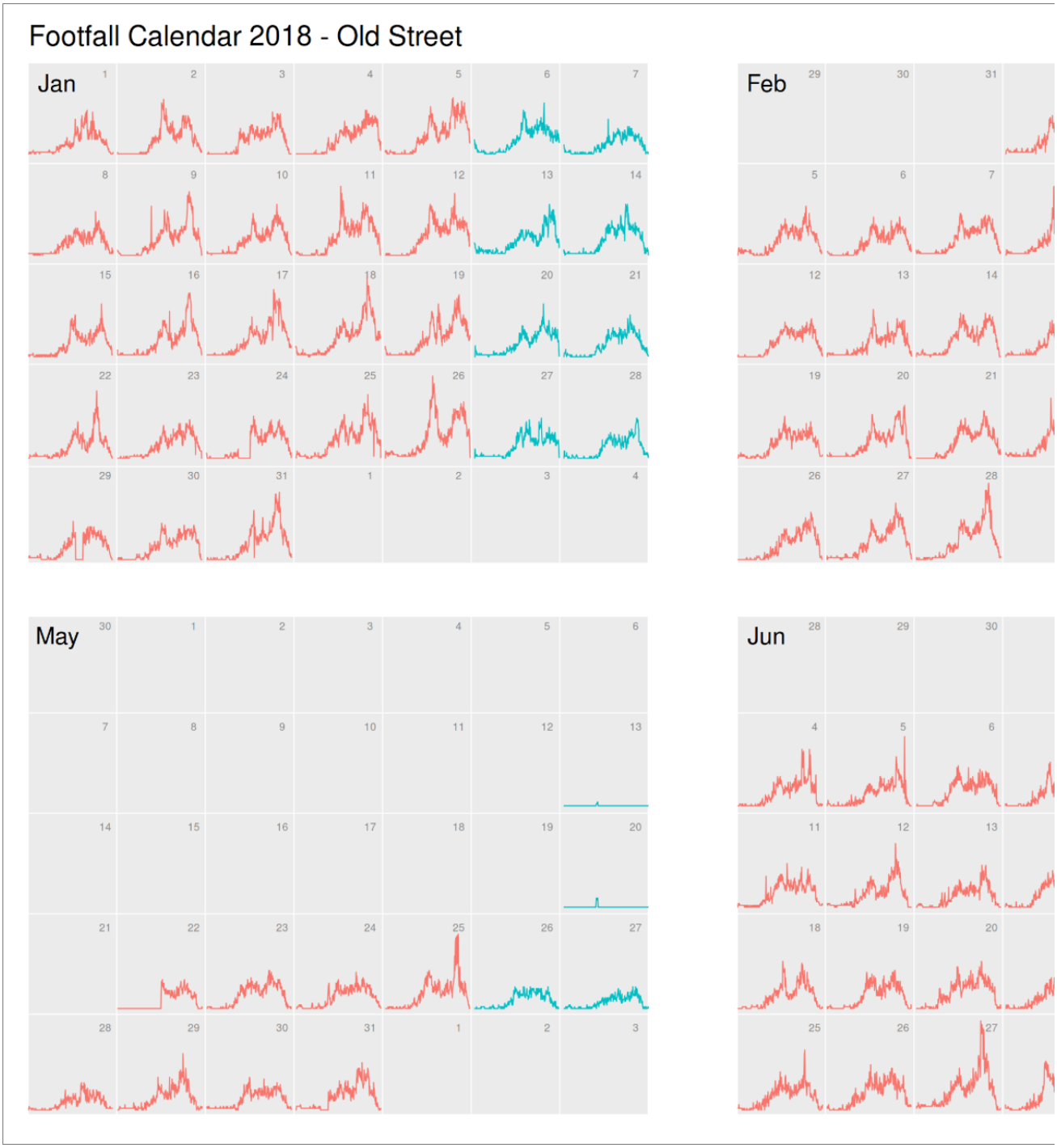
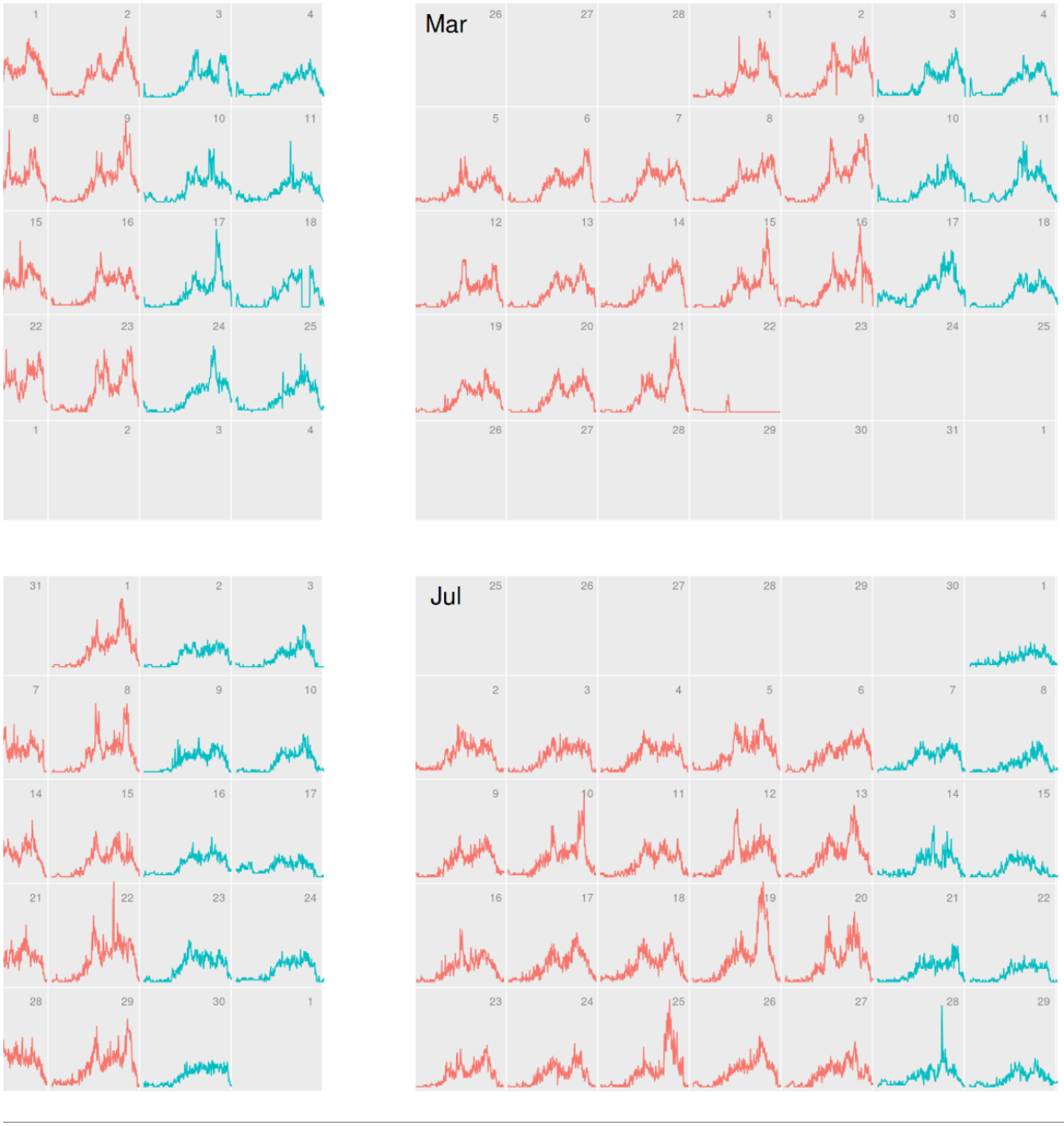


Figure 5.6: Footfall calendar showing the profiles of daily volumes of footfall at Old Street, London.


Mar
Jul

5.2 *Event Detection*

When footfall is looked at longitudinally across locations, a wide range of information can be uncovered about the context which resulted in the patterns in the footfall. Figure 5.7 shows the normalised weekly footfall of 10 different locations across Cardiff for two years: 2017 and 2018. The patterns in the footfall clearly show numerous events that were happening in Cardiff as unusually high or low footfall in the corresponding week. The most significant event was in February 2018, when all sensors reported the lowest numbers they have ever recorded. This coincided with the cold wave in UK named 'Beast from the East', which brought adverse weather conditions all over the UK and led to a significant reduction in footfall. The other identifiable events are bank holiday weekends which result in higher than normal footfall, and the Christmas shopping season when footfall is at its highest. Finally, it is interesting to see the difference in summertime footfall between 2017 and 2018, which could be explained by the FIFA World Cup which took place in the summer of 2018. This example shows the usefulness of the footfall data to detect real life events from the data in near real time. It can also be used to measure the effect of events on footfall, and hence understand the impact of these events for retail and the economy more generally.

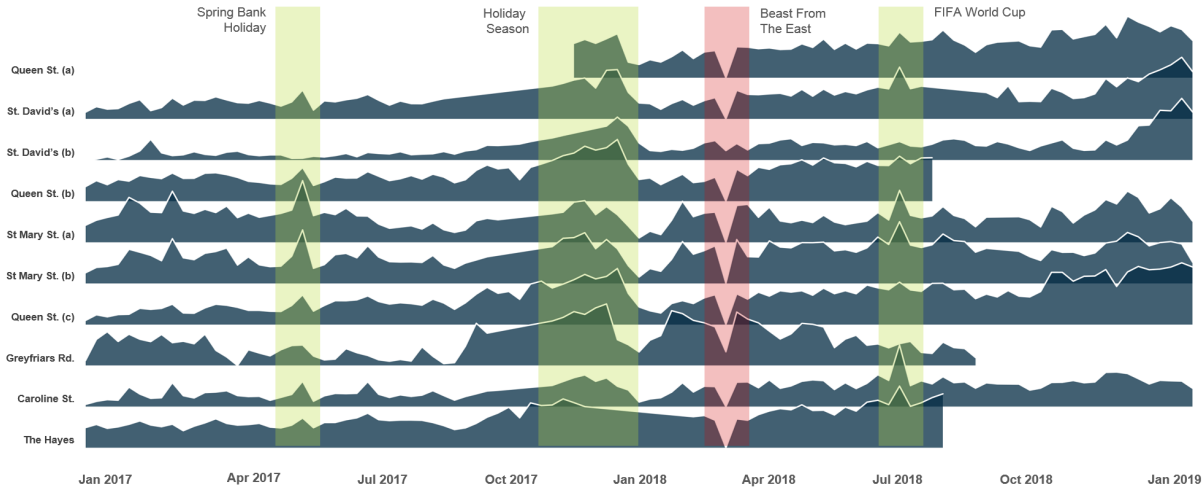


Figure 5.7: Normalised weekly footfall index at locations across Cardiff from 2017 to 2018

5.2.1 *Football world cup*

In addition to long-term changes and events, the footfall data can be used to identify the smaller effects of these events at an area scale. Figure 5.8 shows footfall from two days in Leicester square, London when the quarterfinal and semifinal matches of the 2018 FIFA World Cup took place. Both matches happened in the evening and led to an increase in footfall around match time. The most interesting observation is the effect the outcome of the game had on footfall. On the day of the quarterfinal, the winning result of the English team led to a post-match celebration which pushed the Leicester Square footfall to its day-time highest, unlike the day of the semifinal when the English team lost. This not only shows the usefulness of the data in understanding the effect events have on local footfall, but it also shows how the data can be used by retailers to predict the effect the results of sports events might have on them.

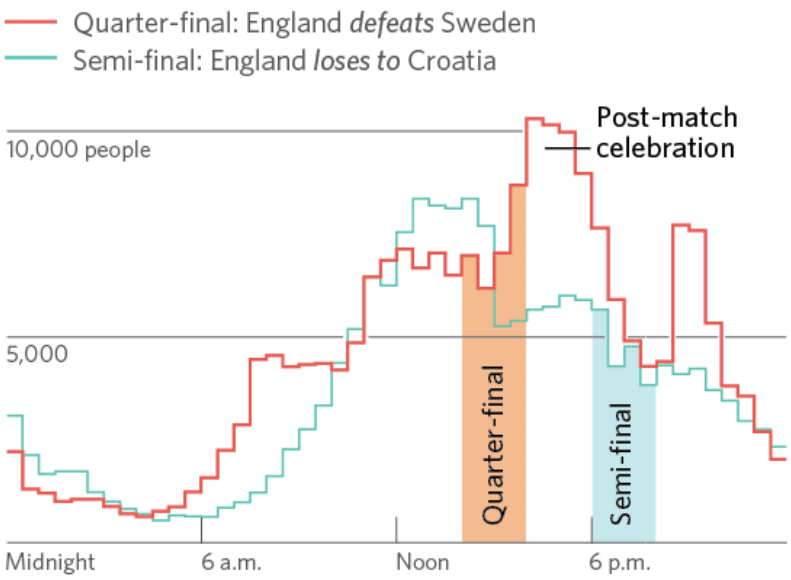


Figure 5.8: The difference in footfall distribution at Leicester square, London after the FIFA World Cup quarterfinal and semifinal matches. Source: Oliver Uberti and James Cheshire

These examples show the importance of footfall data in detecting events. Even a simple visual analytics of the dataset reveal interesting information on events. This would be much more useful when used in tandem with advanced machine learning/data mining techniques, and will predict much better results as more data is collected.

5.3 *Pedestrian Flows*

Detecting general trends in the flow of people between spatial locations is neither obvious nor a trivial task. This is due the high cost of cap-

turing these movements without compromising people's privacy, since the primary way to collect such detailed data involves handling people's precise location data. This research specifically removes any personally identifiable information because of MAC randomisation and hashing, and therefore seems like it might not be suitable for studies on human mobility. However, this problem can be solved by examining the movement of people in the Smart Street Sensors network at a fine spatial and temporal resolution using a novel methodology in the field of big data which uses mathematical models from information theory: Transfer Entropy (TE). Using an area in central London, this section serves as a case study to demonstrate the usefulness of TE as a measure of the flow of pedestrians[1].

[1] Work undertaken was in collaboration with Roberto Murcio and Karlo Lugomer. The methodology was formulated by Murcio; this author worked on the implementation of the method in the case study.

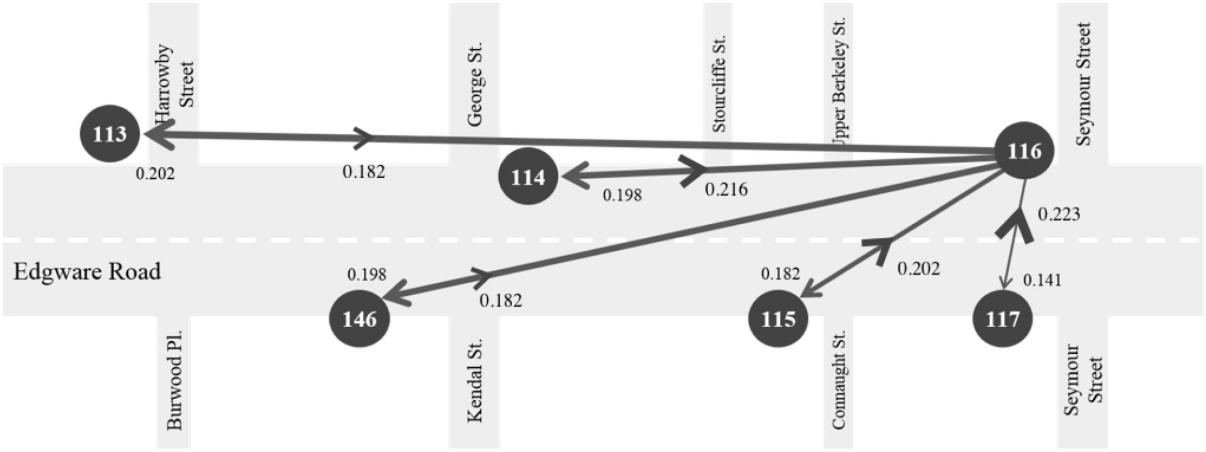


Figure 5.9: Illustration of transfer entropies between set of locations along Edgware Road, London.

Consider the array of sensors shown in Figure 5.9 and assume that there is a flow of people walking past Location 116 and then diffusing towards the remaining sensors. Counts generated by the sensor are aggregated per five minute intervals, so if, for example, it takes one minute to walk from Location 116 to Location 117, the number of people detected at 117 from minutes 2 to 5, would correspond to the percentage of people detected at 116 from minutes 1 to 4. In other words, the similarity between the time series of counts at the locations under consideration are correlated. Hence the aim here is to provide a measure for the size of the flow between each pair of sensors without actually tracking individual people. One way to accomplish this, is to think of this motion of people as flows of information among distinctive sources, so we can relate the number of people reaching one sensor from another by measuring the uncertainty between two interacting random variables. For this, we used

an information theory concept known as Transfer Entropy (TE) defined by:

$$TE(X,Y) = \sum_{t=1} p(y_{t+1}, y_t, x_t) \times log \frac{p(y_{t+1} \mid y_t, x_t)}{p(y_{t+1} \mid y_t)} \tag{5.1}$$

Where t indicates a given point in time. Equation 5.1 measures the reduction in uncertainty at y_t, given x_t and y_{t-1}. In comparison with the case when only y_{t-1} is known. This measure is applied directly to our people's movement problem and X = location i, Y = location j and t runs for a whole day, the TE would represent an indicator for the direction of the flow, as the counts at y_{t+1} are more accurately estimated using the information of x_t.

Taking again Figure 5.9 as a reference, we measured the TE between sensor 116 and the rest of the sensors. The walking time is not constant and each sensor has counts at all times i, j. There are people passing by these sensors that came from locations outside the network. The numbers at each line represent the TE measured between each pair of sensor locations. The largest TE values found were from 114,115 and 117 to 116. The asymmetry of the TE is clear here, as the values in the opposite direction (from 116 to 114,115 and 117) are considerably lower. Another interesting value is the pair 116-117, where TE(116,117) « TE(117,116). This demonstrates that in this four-way crossing, the predominant direction of flow is from Location 117 to Location 116 (from the bottom of the figure upwards, or from west to east in reality). These results suggest that, in general, there is a larger flow of people from the west side to the east side of Edgware Road, and a larger flow of people from south to north of it. The results are consistent with our intuition that there is a larger flow of people from south to north along this road towards Edgware Road underground station.

There is still a series of uncertainties yet to be addressed by this model, such as the decay of probabilities with distance and the number of interventions of opportunity encountered by people while walking from one sensor to another. However, this first initial set of results is encouraging in measuring flow between spatial points without actually tracking these users.

6

Discussion and Conclusions

In the past 30 years there have been an immense change in the way data concerting distribution and dynamics of human population are generated, collected and consumed. Rather than being a top-down, structured designed endeavour, data generation and collection has become a bottom-up procedure where data were created as a result of day to day activities of people and are collected, cleaned, and aggregated into information. There have been a significant volume of research on identifying such data sources and use them for various purposes in both academia and industry. As these data sources got more distributed and comprehensive, the concern to protect user privacy increased as well. This thesis aims to work in between these two the areas of research and the corresponding opportunities present in them. This thesis identifies Wi-Fi probe requests as a source of data from which information on ambient population and behaviour - especially footfall - could be extracted, and solves the problem of inferring accurate footfall information without using personally identifiable information of the users.

6.1 Summary of Findings

From the systematic literature review of around 350 academic publications, it was concluded that Wi-Fi is the most suitable candidate for the technology that can be used to collect data on human presence and movement at a national level. It was found to be the scalable, cheap, universal, and easy way to collect large amounts of granular such data without depending on any other infrastructure. The only shortcomings of projects using Wi-Fi technology are their inherent uncertainties and the leakage of MAC addresses - a globally unique, personally identifiable information which could be related back to the users relatively easily. Finally, two potential areas of research with opportunities for further

study were identified: Creating a standardised, cleaned, accurate and reliable footfall or ambient population from the Wi-Fi probe requests, and solving the specific issue of MAC address randomisation while cleaning and filtering Wi-Fi probe requests. After studying the Wi-Fi specification to get a overall outline of the structure of Wi-Fi probe request, a set of small initial experiments were designed and executed to know more about them. It was found that the number of probe requests and the unique MAC addresses collected by a Wi-Fi sensor is far greater than the number of mobile devices in the immediate vicinity. The signal strength and sequence numbers are some of the important information that are present in the probe requests. Alongside these experiments a longer pilot study was also conducted to result in three distinct datasets. The small experiments collected in-depth data on probe requests at small areas for short intervals. The pilot study covered 5 locations across central London collected data for over two and a half months. Finally the Smart Street Sensor project which collected small set of data from probe requests at 1000 locations continuously for over 3.5 years.

Before moving to cleaning and processing the data, this research undertook a comprehensive look at the nature of these datasets within the context of 'Big data and Big data tools' so that a framework for evaluating the 'bigness' of the datasets can be devised. With such a framework the Wi-Fi datasets were examined in all of their dimensions and found to be 'Medium data' at best. A review of big data tools was also carried out and the tools suitable for the Wi-Fi dataset were picked out and combined together to form a bespoke 'Medium data toolkit' for processing the Wi-Fi data as efficiently as possible.

From the initial exploration, the major uncertainties identified in the data which needs to be solved by the cleaning and processing procedures are, range of the Wi-Fi sensor, differing frequency at which mobile devices emit probe requests, MAC address randomisation which masks the devices unique identification, increasing mobile devices ownership in the population over long-term, missing data from the failures in the sensors and collisions of MAC addresses when they are anonymised using cryptographic hashing. The collisions in hashed MAC addresses were found to be rare and inconsequential.The uncertainty regarding the range of the Wi-Fi sensor was found to cause noise in the data from outside the field of measurement and was solved by filtering out probe requests with low signal strength. This definition of 'Low' signal strength could be deduced dynamically for each location at each time interval

using one dimensional clustering algorithms. The 'k-means' algorithm was found to be best suited for this purpose. The randomisation of MAC addresses lead to over-estimation of number of devices from set of probe requests while the uneven frequency of probe requests emitted by the mobile devices prohibit a simple universal factor for converting number of probe requests to number of devices. It was found that this uncertainty can be solved using a novel graph based methodology which uses the sequence numbers in the packets rather than the MAC address to uniquely identify the devices. When the sequence number is not available the uncertainty can be reduced for an interval by looking at the ratio of number of probe requests to the number of mobile devices in the probe requests without randomised addresses in that interval. These methods along with manual calibration were found to reduce the error in the estimation of footfall from Wi-Fi probe requests to almost 10% at locations with ideal conditions.

In addition to the above data cleaning techniques other processing were done to the probe requests dataset to remove further uncertainties. The missing data in the dataset could be interpolated using a Kalman smoothing based method for short term or a seasonally decomposed method for long term. Finally, the increasing mobile ownership was adjusted using manual counts for short term intervals and using a weekly adjustment factor 0.2% for long term. Using all these methods for filtering, cleaning and adjusting Wi-Fi data, this thesis finalises a overall data processing pipeline for producing a clean, precise, accurate and continuous data on footfall across retail locations across the UK. Finally this research also provides a gallery of examples showing the possible use of such granular and continuous data on footfall on a national level.

6.2 Research Question

Looking back at our research question - "Can dynamics of footfall inferred from passively collected big dataset without putting the privacy of users at risk?", we have demonstrated that the task is indeed feasible, using Wi-Fi probe requests. Even when the identity of the devices were masked using randomisation techniques we have demonstrated that aggregation and estimation could be done without compromising the privacy of the users. In addition to this, we have also demonstrated the usefulness and application of such footfall estimate with various examples. The footfall estimates derived from the method were used to devise a 'footfall index'

at various levels - national, city, area and micro site locations showing how the retail related footfall have been distributed in the UK and how this distribution has been changing over time in high granularity both spatially and temporally. It has been demonstrated that this information on footfall can be used as a clue for knowing the form and function of a place and trace the changes it has undergone over time as well. It was also demonstrated two sets of examples, that real-world events could be detected from looking at the anomalies in the footfall volumes at locations. Finally, it was also demonstrated that such detailed and continuous footfall volume information at locations could be used to predict or estimate flow of pedestrians between them by just looking at the changes in these volumes thus providing a way to understand the pedestrian flow in cities without actually tracking individuals.

6.3 Further Work

As we discussed in the literature review, the research on collecting and using data on population distribution and dynamics have closely followed the advances and changes in the consumer technology. Every new technology adopted for mainstream use spurred new wave of research in using those technology. It is also noted that every new technology not only brought many advantages over the previous ones but also introduced unique challenges. In this context, the larges opportunity in furthering the research exists in identifying, evaluating and adopting new technologies. There is a significant opportunities in applying these new technologies for old challenges and device methods to make them suitable to answer the questions raised by research. Few such technologies are detailed below,

- **5G** is the new generation of technology which aims to bring even higher speeds of data transfer to mobile devices through cellular networks. This may lead to the gradual decline and phasing out of Wi-Fi technology. Though this cellular based technology doesn't provide the similar detail and flexibility offered by Wi-Fi it has the potential to offer much more comprehensive picture of the world if it gets widely adopted.

- **Bluetooth Low energy (BLE)** is the upcoming short-range, wireless personal area network technology. With emphasis on being the technology used by the Internet of Things (IOT) devices, this technology has the potential to displace Wi-Fi as the choice of short-range com-

munications. The explosion of wearables and smart devices at home, the amount of data that could be available form this technology could be staggering in the next decade.

- **Ultra wide band radar** is another short-range technology which has been developed for motion and object detection. Being primarily used to design sensors for proximity and motion detection, this has the potential to become a standard for vehicles. Moreover, with the recent uptick in self driving car research and development, the cost of these devices has gone significantly down thus providing amazing opportunities in creating comprehensive sensor networks similar to Smart Street Sensor project.

In spite of being developed since 1980s, machine learning techniques have received extraordinary interest in the last decade. This interest, along with advancements in the Big data tools and technologies has set up the stage for research by applying supervised and unsupervised machine learning techniques on large scale datasets collected through the above mentioned technologies. There is a significant opportunity for applying unsupervised learning techniques such as anomaly detection and neural networks in passively collected digital data to improve data cleaning, interpolation, population estimations and time series based predictions etc.

Research ethics, safety and privacy are going to be the next big areas of concern for advanced machine learning based techniques and big data analysis in the next decade. The era of uninhibited large scale production, collection and consumption of personal data through connected devices over internet without oversight is almost over. People are increasingly concerned with protecting their privacy and are opposed to the exploitation of their personal data. This concern has been addressed by legislation such as GDPR and technologies such as cryptography and randomisation. All these developments provide us with various opportunities in further research.

Firstly there is opportunity study the above mentioned technologies form a privacy point of view to evaluate the advantages and risks presented by them and advance the research in terms of both mitigation the risks while maintaining some kind of usefulness. These inquiries can not only be done in terms of techniques but also on the lines of legal compliance of such techniques. There is also opportunity for researching on the uncertainties and limits of datasets when subject to robust privacy control methods. Secondly the immense research, innovation

and advancements made in peer to peer technologies in solving the various trust problems could be applied in the field of sensor based population estimation or pedestrian flow detection. There is an opportunity for research into building a peer to peer network of sensors where the data collected by the sensors never leave the device themselves but the analyses are taken to the source of data. This act of "moving the analysis to data" can solve numerous problems of safety of the personal data since there is not central point of failure and it can also scale up indefinitely without overwhelming a central repository of data. Through these further research, we could take the field forward by not only following the improvements in the technology of data collection but also push the envelope in terms of developing more ethical and sage research environment while handling large amounts of data.

7
Appendix

Note : All code used in this thesis has been made available online via github - `https://github.com/sbmkvp/phd-thesis`.

7.1 Manual Counting

7.1.1 Node.js App

```
{
  "name": "manualcount",
  "version": "1.0.0",
  "description": "Manual counting software",
  "main": "manualcount.js",
  "dependencies": {
    "keypress": "^0.2.1",
    "moment": "^2.20.0"
  },
  "devDependencies": {},
  "scripts": {
    "test": "echo \"Error: no test specified\" && exit 1"
  },
  "author": "",
  "license": "ISC"
}
```

```
var keypress = require('keypress');
var moment = require('moment');
keypress(process.stdin);

// listen for the "keypress" event
process.stdin.on('keypress', function (ch, key) {
        console.log('"'+moment().format('YYYY-MM-DD H:mm:s.SSS')+'","'+"1"+'"');
        if (key && key.ctrl && key.name == 'c') { process.stdin.pause(); }
});

process.stdin.setRawMode(true);
process.stdin.resume();
```

7.1.2 Android App

The Android manifest which defines the whole application along with the permissions it needs to function.

```
<?xml version="1.0" encoding="utf-8"?>
<manifest xmlns:android="http://schemas.android.com/apk/res/android"
```

```
  package="com.bala.manualcount"
  android:versionCode="3"
  android:versionName="3">
<uses-sdk android:minSdkVersion="21"/>
<uses-permission android:name="android.permission.WRITE_EXTERNAL_STORAGE"
<uses-permission android:name="android.permission.READ_EXTERNAL_STORAGE" /
<application android:label="Clicker"
  android:icon="@mipmap/counter">
  <activity android:name=".MainActivity">
    <intent-filter>
        <action android:name="android.intent.action.MAIN"/>
        <category android:name="android.intent.category.LAUNCHER"/>
    </intent-filter>
  </activity>
</application>
</manifest>
```

The layout of the app.

```
<?xml version="1.0" encoding="utf-8"?>
<LinearLayout
xmlns:android="http://schemas.android.com/apk/res/android"
  android:layout_width="match_parent"
  android:layout_height="match_parent"
  android:gravity="center"
  android:orientation="vertical"
  android:clickable="true"
  android:focusable="true"
  android:focusableInTouchMode="true"
  android:weightSum="3"
  android:id="@+id/lay">
<TextView
  android:layout_width="fill_parent"
  android:layout_height="fill_parent"
  android:id="@+id/my_text"
  android:textSize="65dp"
  android:layout_weight="1"
  android:gravity="center"/>
<LinearLayout
  android:id="@+id/bottom_layout"
```

```
  android:layout_width="fill_parent"
  android:layout_height="fill_parent"
  android:layout_weight="2"
  android:weightSum="2"
  android:gravity="left|center"
  android:orientation="horizontal" >
<Button
  android:id="@+id/b_left"
  android:layout_width="fill_parent"
  android:layout_height="fill_parent"
  android:layout_weight="1"
  android:text="0"
  style="?android:attr/borderlessButtonStyle"
  android:textSize="20sp" />
<Button
  android:id="@+id/b_right"
  android:layout_width="fill_parent"
  android:layout_height="fill_parent"
  android:layout_weight="1"
  android:text="0"
  style="?android:attr/borderlessButtonStyle"
  android:textSize="20sp" />
</LinearLayout>
</LinearLayout>
```

The main logic of the app.

```
package com.bala.manualcount;

import android.app.Activity;
import android.app.ActionBar;
import android.content.Context;
import android.content.Intent;
import android.view.View;
import android.os.Bundle;
import android.os.Environment;
import android.widget.TextView;
import android.widget.LinearLayout;
import android.widget.Toast;
import android.widget.Button;
```

```
14 import android.net.Uri;
15 import java.util.Random;
16 import java.util.Calendar;
17 import java.text.DateFormat;
18 import java.text.SimpleDateFormat;
19 import java.util.Date;
20 import java.io.File;
21 import java.io.IOException;
22 import java.io.FileOutputStream;
23 import java.io.FileWriter;
24 import java.io.BufferedWriter;
25 import java.io.OutputStreamWriter;
26
27 public class MainActivity extends Activity {
28 @Override
29 protected void onCreate(Bundle savedInstanceState) {
30 ActionBar actionBar = getActionBar();
31 actionBar.hide();
32   super.onCreate(savedInstanceState);
33   setContentView(R.layout.activity_main);
34   TextView text = (TextView)findViewById(R.id.my_text);
35 text.setText("0");
36 File root = new File(Environment.getExternalStorageDirectory(),
37     "ManualCounts");
38 if (!root.exists()) { root.mkdirs(); }
39 Button left = (Button)findViewById(R.id.b_left);
40 left.setOnClickListener(new View.OnClickListener(){
41   @Override
42   public void onClick(View view) { try {
43     Date date = Calendar.getInstance().getTime();
44     DateFormat preciseTime =
45        new SimpleDateFormat("yyyy-MM-dd hh:mm:ss.SSS");
46     DateFormat justDate = new SimpleDateFormat("yyyy-MM-dd");
47     File root = new File(Environment.getExternalStorageDirectory(),
48         "ManualCounts");
49     File csvfile = new File(root,justDate.format(date)+".csv");
50     FileWriter fwriter = new FileWriter(csvfile,true);
51     BufferedWriter writer = new BufferedWriter(fwriter);
52     String strDate = preciseTime.format(date);
```

```
53      writer.append("\""+strDate+"\",\"left\"");
54      writer.newLine();
55      writer.close();
56          TextView text = (TextView)findViewById(R.id.my_text);
57      int count = Integer.parseInt(text.getText().toString());
58      text.setText(Integer.toString(count+1));
59      Button left = (Button)findViewById(R.id.b_left);
60      int countleft = Integer.parseInt(left.getText().toString());
61      left.setText(Integer.toString(countleft+1));
62      } catch (IOException e) {
63      e.printStackTrace();
64      Context context = getApplicationContext();
65      Toast.makeText(context, "error", Toast.LENGTH_SHORT).show();
66      } } });
67  Button right = (Button)findViewById(R.id.b_right);
68  right.setOnClickListener(new View.OnClickListener(){
69    @Override
70    public void onClick(View view) { try {
71      Date date = Calendar.getInstance().getTime();
72      DateFormat preciseTime =
73        new SimpleDateFormat("yyyy-MM-dd hh:mm:ss.SSS");
74      DateFormat justDate = new SimpleDateFormat("yyyy-MM-dd");
75      File root = new File(Environment.getExternalStorageDirectory(),
76          "ManualCounts");
77      File csvfile = new File(root,justDate.format(date)+".csv");
78      FileWriter fwriter = new FileWriter(csvfile,true);
79      BufferedWriter writer = new BufferedWriter(fwriter);
80      String strDate = preciseTime.format(date);
81      writer.append("\""+strDate+"\",\"right\"");
82      writer.newLine();
83      writer.close();
84          TextView text = (TextView)findViewById(R.id.my_text);
85      int count = Integer.parseInt(text.getText().toString());
86      text.setText(Integer.toString(count+1));
87      Button right = (Button)findViewById(R.id.b_right);
88      int countright = Integer.parseInt(right.getText().toString());
89      right.setText(Integer.toString(countright+1));
90      } catch (IOException e) {
91      e.printStackTrace();
```

```
92      Context context = getApplicationContext();
93      Toast.makeText(context, "error", Toast.LENGTH_SHORT).show();
94      } } });
95  LinearLayout layOut = (LinearLayout)findViewById(R.id.lay);
96  layOut.setOnClickListener(new View.OnClickListener(){
97    @Override
98    public void onClick(View view) { try {
99      Date date = Calendar.getInstance().getTime();
100     DateFormat preciseTime =
101       new SimpleDateFormat("yyyy-MM-dd hh:mm:ss.SSS");
102     DateFormat justDate = new SimpleDateFormat("yyyy-MM-dd");
103     File root = new File(Environment.getExternalStorageDirectory(),
104         "ManualCounts");
105     File csvfile = new File(root,justDate.format(date)+".csv");
106     FileWriter fwriter = new FileWriter(csvfile,true);
107     BufferedWriter writer = new BufferedWriter(fwriter);
108     String strDate = preciseTime.format(date);
109     writer.append("\""+strDate+"\",\"other\"");
110     writer.newLine();
111     writer.close();
112         TextView text = (TextView)findViewById(R.id.my_text);
113     int count = Integer.parseInt(text.getText().toString());
114     text.setText(Integer.toString(count+1));
115     } catch (IOException e) {
116     e.printStackTrace();
117     Context context = getApplicationContext();
118     Toast.makeText(context, "error", Toast.LENGTH_SHORT).show();
119     } } });
120 layOut.setOnLongClickListener(new View.OnLongClickListener(){
121   @Override
122   public boolean  onLongClick(View view) {
123     Context context = getApplicationContext();
124     Toast.makeText(context,"Sending data",Toast.LENGTH_SHORT).show();
125     Intent intentShareFile = new Intent(Intent.ACTION_SEND);
126     Date date = Calendar.getInstance().getTime();
127     DateFormat justDate = new SimpleDateFormat("yyyy-MM-dd");
128     File root = new File(Environment.getExternalStorageDirectory(),
129         "ManualCounts");
130     File csvfile = new File(root,justDate.format(date)+".csv");
```

```
131     intentShareFile.setType("text/*");
132     intentShareFile.putExtra(Intent.EXTRA_STREAM,
133         Uri.parse("file://"+csvfile.toString()));
134     startActivity(Intent.createChooser(intentShareFile,
135           "Share File"));
136     return true;
137   }
138 }); } }
```

7.2 Pilot Study

7.2.1 Sensor

```
{
  "name": "smart_street_sensor",
  "version": "1.0.0",
  "description": "Tools for collecting data",
  "main": "server.js",
  "scripts": {
    "start": "node server.js"
  },
  "keywords": [
    "footfall",
    "data",
    "collection",
    "wifi",
    "probe",
    "requests"
  ],
  "author": "Balamurugan Soundararaj",
  "license": "GNU-GPL",
  "dependencies": {
    "adler32": "^0.1.7",
    "csv": "^2.0.0",
    "md5": "^2.2.1",
    "moment": "^2.20.1",
    "node-schedule": "^1.3.0",
    "socket.io": "^2.0.4",
    "socket.io-client": "^2.0.4",
    "stream": "0.0.2",
    "util": "^0.10.3",
    "zlib": "^1.0.5"
  }
}
```

```
// ================================================
// Import the required modules
// ================================================
const csv = require('csv');
```

```
5  const crypto = require('crypto');
6  const adler32 = require('adler32');
7  const Writable = require("stream").Writable;
8  const moment = require('moment');
9  const schedule = require('node-schedule');
10 const zlib = require('zlib');
11 const io = require('socket.io-client')
12
13 // ------------------------------------------------
14 // Global variables to buffer the output for every
15 // 5 minutes, server ip address
16 // and port where the data needs to be pushed and
17 // the sensor id which is sending
18 // the data.
19 // ------------------------------------------------
20 var timestamp = '';
21 var buffer = [];
22 var server_address = process.argv[2];
23 var sensor_id = process.argv[3];
24
25 // ================================================
26 // Create a new websocket connection and on
27 // connection to the server.
28 // ================================================
29 var socket = new io('http://'+server_address);
30
31 // ================================================
32 // Scheduler which is invoked every 5 minutes
33 // which sends the buffered data in
34 // the global variable to the server by emitting
35 // event in the socket
36 // ================================================
37 schedule.scheduleJob('00 */5 * * * * *', function(){
38   if(timestamp!='') {
39     var data = {
40       sensor : sensor_id,
41       timestamp : timestamp,
42       data : buffer
43     }
```

```
    timestamp = '';
    buffer = [];
    data = zlib.gzipSync(JSON.stringify(data));
    socket.emit('data',data);
  }
});

// ================================================
// create a writable steam which buffers the probe
// requests collected into the
// global data variable. Which is flushed and sent
// to the server at regular schedules.
// ================================================
var buffer_data = Writable({objectMode:true});
buffer_data._write = function (chunk, encoding, next) {
  timestamp = chunk.splice(6,1);
  buffer.push(chunk);
  next();
}

// ================================================
// Modify the read data - Split the MAC address
// into two parts, hash the user
// part using MD5 algorithm and format the date
// field better.
// ================================================
function clean_record(record){
  d_form = 'MMM DD, YYYY HH:mm:ss.SSSSSSS';
  var mac_split = record[1].split(":");
  var oui = mac_split[0]+
    mac_split[1]+
    mac_split[2];
  record[1] = crypto.createHash('md5')
    .update(record[1])
    .digest('hex')
    .toString();
  record[1] = adler32.Hash().
    update(record[1]).digest('hex')
  record.push(oui);
```

```
  record.splice(1,0,moment(record[0], d_form)
    .format('mm:ss.SSS'));
  record.push(moment(record[0], d_form)
    .format('YYYY-MM-DD HH:'));
  record.splice(0,1);
  return(record)
}

// ================================================
// Take the input from stdin and pipe it through
// the series of functions we setup earlier to
// emit a stream of http post requests.
// ================================================
process
  .stdin
  .pipe(csv.parse())
  .pipe(csv.transform((record) =>
    {return(clean_record(record));}))
  .pipe(buffer_data)
```

```
#!/bin/sh
sudo tshark -Iql \
   -i $1 \
   -T fields \
   -E separator=, \
   -E quote=d \
   -e frame.time \
   -e wlan.sa \
   -e wlan_radio.signal_dbm \
   -e frame.len \
   -e wlan.seq \
   type mgt subtype probe-req and broadcast
```

7.2.2 *Server*

```
{
  "name": "server",
  "version": "1.0.0",
  "description": "Recieving and storing the data",
  "main": "server.js",
```

```
  "author": "Balamurugan Soundararaj",
  "license": "ISC",
  "dependencies": {
    "csv": "^2.0.0",
    "moment": "^2.20.1",
    "object-sizeof": "^1.2.0",
    "socket.io": "^2.0.4",
    "zlib": "^1.0.5"
  }
}
```

```
// ================================================
// Import the required modules
// ================================================
const http = require('http').Server();
const { exec } = require('child_process');
const io = require('socket.io')(http);
const zlib = require('zlib');
const sizeof = require('object-sizeof');
const moment = require('moment');

// ================================================
// The function which takes the data received and
// pushes it to the database.
// ================================================

function store_data(data) {
  var cmd = 'echo "'+data+
    '" | psql sss -U '+process.argv[3]+
    ' -c "copy probes from stdin with delimiter \',\';"';
  exec(cmd,(err,stdout,stderr)=>{});
}

// ================================================
// Function to convert the JSON data received into
// csv string.
// ================================================

function format_data(data){
  data = zlib.gunzipSync(data).toString();
```

```
30    data = JSON.parse(data);
31    var csv_string =''
32    for(var i=0; i < data.data.length; i++) {
33      var line = ''
34      data.data[i][0]=data.timestamp+data.data[i][0];
35      data.data[i][6]=data.sensor;
36      for(j in data.data[i]) {
37        line = line+'"'+data.data[i][j]+'"';
38        if(j < data.data[i].length-1){ line = line+','; }
39      }
40      csv_string = csv_string+'\n'+line
41      if(i%1000 == 0 || i >= data.data.length-1){
42        store_data(csv_string.trim());
43        csv_string = '';
44      }
45  }
46  console.log(moment().
47    format("YYYY-MM-YY HH:mm")+" - "+data.sensor);
48  }
49
50  // ================================================
51  // Setting up the server to use a text parser and
52  // configure the route to execute the store_data
53  // function when it receives the data from the sensor
54  // ================================================
55  io.on('connection',function(socket){
56    socket.on('data', function(data){
57      process.stdout.write((sizeof(data)/1024)
58        .toFixed(2)+" - ");
59      format_data(data);
60    });
61  });
62
63  // ================================================
64  // The app listens at the port specified as first
65  // commanline argument
66  // ================================================
67  http.listen(process.argv[2]);
```

7.3 Data Pipeline

7.3.1 Sample configuration file

```
{
  "sas_api": "url_string",
  "sas_token": "token_string",
  "ldc_server": "url_to_the_ldc_database",
  "ldc_user": "username_for_ldc_database",
  "ldc_pass": "password_for_ldc_database",
  "pg_user": "user_name_for_database",
  "pg_pass": "user_password_for_database",
  "vendors": "https://code.wireshark.org/...",
  "raw": "../path/to/raw_data/folder",
  "hashed": "../path/to/probe_requests/folder",
  "encrypted": "../path/to/encrypted/folder",
  "salt": "random_string_here"
}
```

7.3.2 Data Pipeline

```
#! /bin/bash

#------------------------------------------------
# Print usage information to stdout for help
#------------------------------------------------
function usage() {
        echo "Usage: pipeline [options]";
        echo "Options:";
        echo -e "  --help      Display this message.";
        echo -e "  --date      Date in the format \
    yyyy/mm/dd. Defaults to today's date";
        echo -e "  --config    The config file to be \
    used. Defaults to ./config.json";
        echo -e "  --meta      Set this to 'false' to \
    not update meta-data";
}

#------------------------------------------------
```

```
# Set default variables for the parameters
#-----------------------------------------------
DATE=`date +'%Y/%m/%d'`
CONFIG="config.json"
META="true"
DOWN="true"
PROCESS="true"

#-----------------------------------------------
# Read in the named parameters from the commandline
#-----------------------------------------------
while [ "$1" != "" ]; do
        PARAM=`echo $1 | awk -F= '{print $1}'`
        VALUE=`echo $1 | awk -F= '{print $2}'`
        case $PARAM in
                --help) usage; exit ;;
                --date) DATE=$VALUE ;;
                --config) CONFIG=$VALUE ;;
                --meta) META=$VALUE ;;
                --down) DOWN=$VALUE ;;
                --process) PROCESS=$VALUE ;;
                *) echo "[$(date +'%Y/%m/%d %H:%M:%S')]: \
      Option \"$PARAM\" is unknown"; exit 1 ;;
        esac
        shift
done

#-----------------------------------------------
# Check the validity of all parameters
#-----------------------------------------------
if [ "$(date +'%Y/%m/%d' -d $DATE 2>/dev/null)"\
  != $DATE ]; then
        echo "[$(date +'%Y/%m/%d %H:%M:%S')]:\
    Date $DATE is invalid"; exit; fi
if [ ! -f $CONFIG ]; then
        echo "[$(date +'%Y/%m/%d %H:%M:%S')]:\
    Cannot find the file $CONFIG"; exit; fi

#-----------------------------------------------
```

```
# Setting the values of all the required variable
# and creating the folders if they don't exist
#-----------------------------------------------
SASAPI=`cat $CONFIG \
        | jq -r '.sas_api'`
SASTOKEN=`cat $CONFIG \
        | jq -r '.sas_token'`
RAWLOC=`cat $CONFIG \
        | jq -r 'if (.raw==null or .raw=="") \
    then "raw" else .raw end'`
HASHLOC=`cat $CONFIG \
        | jq -r 'if (.hashed==null or .hashed=="") \
    then "probes" else .hashed end'`
CRYPTLOC=`cat $CONFIG \
        | jq -r 'if (.encrypted==null or .encrypted=="") \
    then "encrypted" else .encrypted end'`
LDCUSER=`cat $CONFIG \
        | jq -r '.ldc_user'`
LDCPASS=`cat $CONFIG \
        | jq -r '.ldc_pass'`
LDCSERVER=`cat $CONFIG \
        | jq -r '.ldc_server'`
VENDORS=`cat $CONFIG \
        | jq -r '.vendors'`
PGUSER=`cat $CONFIG \
        | jq -r '.pg_user'`
PGPASS=`cat $CONFIG \
        | jq -r '.pg_pass'`
SALT=`cat $CONFIG \
        | jq -r '.salt'`

#-----------------------------------------------
# Setting up the environment variables
#-----------------------------------------------
export PGPASSWORD=$PGPASS
RAWLOC=$RAWLOC/$DATE
HASHLOC=$HASHLOC/$DATE
CRYPTDIR="$CRYPTLOC/$(echo $DATE | awk -F'/' \
  '{print $1"/"$2}')"
```

```
CRYPTFILE="$CRYPTDIR/$(echo $DATE | awk -F'/' \
  '{print $3}').zip"
if [ ! -d $RAWLOC ]; then mkdir -p $RAWLOC; fi
if [ ! -d $HASHLOC ]; then mkdir -p $HASHLOC; fi
if [ ! -d $CRYPTDIR ]; then mkdir -p $CRYPTDIR; fi

#-----------------------------------------------
# Setting up the commands to execute
#-----------------------------------------------
LOG="date +%Y/%m/%d-%H:%M:%S"
HASH="./scripts/flatten $RAWLOC/{} \
        | ./scripts/hash $SALT > $HASHLOC/{}.csv"
LOCATE="./scripts/locate \"\$(echo {} \
        | awk -F'.' '{print \$1}')\" $DATE $PGUSER \
        | awk '{print \$0} END{if(NR==0) print \"0\"}'"
COUNT="cat $HASHLOC/{} \
        | ./scripts/count \`$LOCATE\` \
        | ./scripts/adjust 2> /dev/null \
        | ./scripts/impute 2> /dev/null \
        | psql -q footfall -U $PGUSER \
    -c \"copy counts from stdin with csv\""

#-----------------------------------------------
# Data pipeline.
# Download > Anonymise > Encrpt > Meta > Count
#-----------------------------------------------
echo "[`$LOG`]: Pipeline started for $DATE." &&

if [ $DOWN != "false" ]; then
# Script to download the data--------------------
{
        {
                ./scripts/download \
      $DATE $RAWLOC $SASAPI $SASTOKEN;
        } || {
                echo "[`$LOG`]: \
      Download failed {$DATE:::$SALT}"; exit;
        }
} &&
```

```
if [ "$(ls $RAWLOC | wc -l)" = "0" ]; then
        echo "[`$LOG`]: \
    No files downloaded!! {$DATE:::$SALT}"; exit;
fi
echo "[`$LOG`]: Download completed for $DATE." &&

# Hash the MAC addresses-------------------------
ls "$RAWLOC" | parallel $HASH &&
echo "[`$LOG`]: Hashing completed for $DATE." &&

# Encrypt the data for transfer-----------------
if [ -f $CRYPTFILE ]; then rm $CRYPTFILE; fi &&
gpg-zip -e -r ucl-team@cdrc.ac.uk\
  -o $CRYPTFILE $RAWLOC 2> /dev/null &&
echo "[`$LOG`]: \
  Encryption completed for $DATE." &&
rm -rf $RAWLOC &&
echo "[`$LOG`]: Raw files deleted for $DATE.";
fi &&

# Download the meta data-------------------------
if [ $META != "false" ]; then
./scripts/meta_data $LDCSERVER $LDCUSER \
  $LDCPASS $VENDORS $PGUSER &&
echo "[`$LOG`]: Meta-data downloaded for $DATE.";
fi &&

#Count the probe requests-----------------------
if [ $PROCESS != "false" ]; then
ls $HASHLOC | parallel "$COUNT" &&
echo "[`$LOG`]: Processing completed for $DATE.";
fi
unset PGPASSWORD
```

7.3.3 *Daily Download*

```
#! /bin/bash

#-------------------------------------------------
```

```
# Assigning variables from positional arguments
#-------------------------------------------------
cd /store2/tools
DATE=`date +'%Y/%m/%d' -d"-1day"`

#-------------------------------------------------
# Execute the pipeline and log outputs
#-------------------------------------------------

./pipeline --date=$DATE \
  --config=./config_daily.json \
        1>> ../logs/daily.log \
        2>> ../logs/daily.error
```

7.3.4 *Batch Download*

```
#! /bin/bash

#-------------------------------------------------
# Assigning variables from positional arguments
#-------------------------------------------------
FROM=$1
TO=`date +'%Y/%m/%d' -d"$2+1day"`
CONFIG=$3
META=$4

#-------------------------------------------------
# Iterate through the date range sequentially and
# apply data processing pipeline. parallel is not
# used since each individual pipeline is
# parallelised
#-------------------------------------------------
while [[ $FROM < $TO ]] ; do
        if [ `date +'%u' -d"$FROM"` == 1 ];
  then ./scripts/rotate_salt $CONFIG; fi
        ./pipeline --date=$FROM --config=$CONFIG
        FROM=`date +'%Y/%m/%d' -d"$FROM+1day"`;
done
```

7.3.5 *Component Scripts*

Download the data from Azure data store.

```
#! /bin/bash

#-----------------------------------------------
# Assigning variables from positional arguments
#-----------------------------------------------
DATE=$1
LOCATION=$2
SASAPI=$3
SASTOKEN=$4

#-----------------------------------------------
# Getting the SAS URI for the azure blob container
#-----------------------------------------------
HEADER="Authorization: accessToken $SASTOKEN"
ACCESSURI=`curl -s --header "$HEADER" $SASAPI \
  | jq -r '.Data'`
SOURCE="$(echo $ACCESSURI \
  | awk -F'?' '{print $1}')"
SOURCESAS="?$(echo $ACCESSURI \
  | awk -F'?' '{print $2}')"

#-----------------------------------------------
# Download the container contents for the
# specified date
#-----------------------------------------------
azcopy \
        --source "$SOURCE/$DATE" \
        --destination "$LOCATION" \
        --source-sas "$SOURCESAS" \
        --recursive \
        --quiet
```

Transform the JSON files to CSV.

```
#! /bin/sh

# -----------------------------------------------
```

```
# Setup the variables from input
# ------------------------------------------------
directory=$1
device=$(echo $1 | awk -F'/' '{print $NF}')
date=$(echo $1 | awk -F'/' \
  '{print $(NF-3)"-"$(NF-2)"-"$(NF-1)}')

# ------------------------------------------------
# Set up jq filter string.
# Convert the individual json to csv
# ------------------------------------------------
jq_string=".[] \
  | [.VendorMacPart+.MacAddress, \
  .Signal, .PacketCount,
        ( (input_filename/\"/\" \
    | .[ .|length-1 ] )/\".\" | .[0]),
        .VendorMacPart] | @csv"

# ------------------------------------------------
# Set up the awk print string.
# Add device and timestamp
# ------------------------------------------------
awk_string="{print \"$date \"substr(\$4,2,2)\":\
  \"substr(\$4,4,2)\":00\",
        \"$device\",\$1,\$5,\$2,\$3}"

# ------------------------------------------------
# Set up the shell command.
# Find all .pd files, run jq on them and run awk
# on the output
# ------------------------------------------------
cmd="jq -r '$jq_string' $directory/*.pd \
  | awk -F, -v OFS=, '$awk_string'";

# ------------------------------------------------
# Execute and echo the command.
# ------------------------------------------------
echo "$(eval $cmd)";
```

Hash the MAC address field.

```
#! /usr/bin/Rscript

#---------------------------------------------
# Load required libraries
#---------------------------------------------
suppressMessages(library('openssl'))
suppressMessages(library('tidyverse'))

#---------------------------------------------
# Get the salt from config file
#---------------------------------------------
salt <- commandArgs(trailingOnly = TRUE)[1]

#---------------------------------------------
# Read from stdin, hash the MAC address using
# SHA256 and write to stdout
#---------------------------------------------
read.table(file('stdin'),
                   header = FALSE,
                   sep = ",",
                   quote = "\"",
                   stringsAsFactors = FALSE) %>%
        mutate(V3 = sha256(paste0(V3,salt))) %>%
        format_csv(col_names = FALSE) %>%
        cat

#---------------------------------------------
```

Find the locations of the sensors.

```
#! /bin/bash

#---------------------------------------------
# Assigning variables from positional arguments
#---------------------------------------------
DEVICE=$1
DATE="$(date +'%Y-%m-%d' -d$2)"
PGUSER=$3

#---------------------------------------------
# Setting up arguments for later use
```

```
#-----------------------------------------------
CONNECT_UCL="psql -q footfall -U $PGUSER"
QUERY="copy (select location from installs where\
  device = $DEVICE and start_date <='$DATE' and\
  (end_date >= '$DATE' or end_date is null))\
  to stdout with csv"

#-----------------------------------------------
# Update all the tables
#-----------------------------------------------
$CONNECT_UCL -c "$QUERY"
```

Aggregate the counts based on MAC addresses.

```
#! /usr/bin/Rscript

# ------------------------------------------------
# Load necessary packages
# ------------------------------------------------
suppressMessages(library('tidyverse'))
suppressMessages(library('lubridate'))

# ------------------------------------------------
# Common functions
# ------------------------------------------------
is_moving <- function(x) {
  return( !(((x - (5 * 60)) %in% x) |
        ((x - (10 * 60)) %in% x) |
        ((x - (15 * 60)) %in% x) |
        ((x - (20 * 60)) %in% x) |
        ((x - (25 * 60)) %in% x) |
        ((x - (30 * 60)) %in% x) ) ) }

filter_moving <- function(x) {
        data_vector <- unlist(x[[1]])
        logical_vector <- unlist(x[[2]])
        return( list(data_vector[logical_vector] )) }

flatten_list <- function(x) {
        return( data.frame(mac = x[[1]],
                timestamp = x[[2]],
```

```
                vendor = x[[3]],
                stringsAsFactors = FALSE) ) }

# --------------------------------------------
# Read data from standard input
# --------------------------------------------
data_in <- read.table(file("stdin"),
                      header = FALSE, sep = ",",
                      quote = "\"",
        stringsAsFactors = FALSE)
names(data_in) <- c("timestamp", "device", "mac",
                  "vendor", "signal", "packets")
data_in <- data_in %>%
        mutate(timestamp = round_date(
    as.POSIXct(timestamp,tz="UTC"),"5 min"),)
this_date <- substr(
    as.character(data_in$timestamp[1]),1,10)
location <- commandArgs(trailingOnly = TRUE)[1]
device <- data_in$device[1]
# --------------------------------------------
# Generate overall counts
# --------------------------------------------
counts_overall <- data_in %>%
        mutate(
                timestamp = as.character(timestamp),
                vendor = (vendor %>%
              substr(2,2) %>%
              tolower()) %in%
    c("e","a","2","6")) %>%
        group_by(timestamp) %>% summarize(
                probes_global = packets[!vendor] %>% sum,
                probes_local = packets[vendor] %>% sum,
                macs_global = mac[!vendor] %>%
      unique %>% length,
                macs_local = mac[vendor] %>%
      unique %>% length) %>%
        data.frame(stringsAsFactors=FALSE)

# --------------------------------------------
```

```
# Filter the dataset and generate filtered counts
# -----------------------------------------------
one_time <- data_in %>%
        group_by(mac) %>%
        filter(n()<2) %>%
        select(mac,timestamp,vendor) %>%
        mutate( timestamp = as.character(timestamp))%>%
        data.frame

if( (data_in %>%
     group_by(mac) %>%
     filter(n()>1) %>% nrow) > 0 ) {
repeated <- data_in %>%
        group_by(mac) %>%
        filter(n()>1) %>% ungroup() %>%
        select(mac,timestamp,vendor) %>%
        group_by(mac) %>%
        summarize(
                timestamps = timestamp %>% list,
                vendors = vendor %>% list) %>%
        mutate(is_moving =
    lapply(timestamps, is_moving)) %>%
        mutate(
                timestamps = apply(.[ ,c(2, 4)],
                       1, filter_moving),
                vendors = apply(.[ , c(3, 4)],
                    1, filter_moving)) %>%
        mutate(
                timestamps = lapply(timestamps,
      function(x){return(x[[1]])}),
                vendors = lapply(vendors,
      function(x){return(x[[1]])})) %>%
        select(-is_moving) %>%
        apply(1, flatten_list) %>%
    do.call("rbind", .) %>%
        mutate( timestamp = as.character(timestamp))
} else {
repeated <- data_in %>%
        group_by(mac) %>%
```

```
        filter(n()>1) %>%
        mutate(timestamp =
         as.character(timestamp)) %>%
        data.frame
}
counts_filtered <- rbind(one_time,repeated) %>%
        mutate(
                vendor = (vendor %>% substr(2,2) %>%
              tolower()) %in%
    c("e","a","2","6")) %>%
        group_by(timestamp) %>% summarize(
                count_global = mac[!vendor] %>%
      unique %>% length,
                count_local = mac[vendor] %>%
      unique %>% length ) %>%
        data.frame

# ------------------------------------------------
# Write data to standout output
# ------------------------------------------------
final_counts <- list(counts_overall,
                     counts_filtered) %>%
  reduce(left_join, by = "timestamp")
start_timestamp <- as.POSIXct(
  paste(this_date,"00:00:00"),tz="UTC")
end_timestamp <- as.POSIXct(
  paste(this_date,"23:55:00"),tz="UTC")
final_counts <- left_join(
        data.frame(
                timestamp = seq(start_timestamp,
                    end_timestamp,
                    by = "5 min") %>%
                        format("%Y-%m-%d %H:%M:%S"),
                stringsAsFactors=FALSE),
        final_counts,
        by="timestamp")
final_counts[is.na(final_counts)] <- 0
final_counts$location <- location
final_counts$device <- device
```

```
final_counts %>%
  select(timestamp, location, device,
         probes_global, probes_local,
         macs_global, macs_local,
         count_global, count_local) %>%
        mutate_if(is.numeric,as.integer) %>%
        mutate_if(is.integer,as.character) %>%
        format_csv(col_names = FALSE) %>% cat
```

Adjust the local counts based on probes to MAC ratio.

```
#! /usr/bin/Rscript

# --------------------------------------------
# Load tidyverse for pipes
# --------------------------------------------
suppressMessages(library('tidyverse'))

# --------------------------------------------
# Read data from the standard input
# --------------------------------------------
data <- read.table(file("stdin"),
    header = FALSE,
    sep = ",",
    quote = "\"",
    stringsAsFactors = FALSE)
names(data) <- c("ts","loc","dev","pg",
    "pl","mg","ml","cg","cl")

# --------------------------------------------
# adjusted value based on the "dwellingness" at
# the location at that interval which is inferred
# from the global counts
# --------------------------------------------
data %>%
        mutate(adj = as.integer(ifelse(pl==0|ml==0|cl==0,
          0,ceiling(ifelse(pg>0,cg/pg*pl,cl))))) %>%
        mutate_if(is.numeric,as.integer) %>%
        mutate_if(is.integer,as.character) %>%
        format_csv(col_names=FALSE) %>% cat
```

Impute missing values using Kalman filter.

```
#! /usr/bin/Rscript

# ------------------------------------------------
# Load tidyverse for pipes and imputeTS for
# imputing methods
# ------------------------------------------------
suppressMessages(library('tidyverse'))
suppressMessages(library('imputeTS'))

# ------------------------------------------------
# Read data from the standard input
# ------------------------------------------------
data <- read.table(file("stdin"), header = FALSE,
   sep = ",", quote = "\"",
   stringsAsFactors = FALSE)
names(data) <- c("ts","loc","dev","M1","M2",
                 "M3","M4","M5","M6","M7")

# ------------------------------------------------
# Find just the gaps shorter than 30 mins
# ------------------------------------------------
find_short <- function(d,n) {
        d <- data.frame(x = d$M1+d$M2,
                  g=cumsum(c(1,
                    diff(d$M1+d$M2) != 0)))
        d <- d %>% group_by(g) %>%
    summarise(c = length(x)) %>%
    left_join(d,.,by="g")
        return(d$x==0 & d$c<=n)}
model <- find_short(data,6)
if(nrow(data[model,])>0) {
  data[model,][,c(4:10)] <- NA }

# ------------------------------------------------
# Impute the NA values in each column using
# kalman method. Hyndman RJ and Khandakar Y
# (2008). "Automatic time series forecasting: the
# forecast package for R".
```

```
# Journal of Statistical Software, 26(3).
# ------------------------------------------------
data <- data %>%
  mutate_at(vars(starts_with("M")),
            funs(as.integer(na.kalman(.))))
data[data<0] <- 0
data$model <- model
data %>%
        mutate_if(is.numeric,as.integer) %>%
        mutate_if(is.integer,as.character) %>%
        format_csv(col_names=FALSE) %>% cat
```

Download and update meta data.

```
#! /bin/bash

#------------------------------------------------
# Assigning variables from positional arguments
#------------------------------------------------
LDCSERVER=$1
LDCUSER=$2
LDCPASS=$3
VENDORS=$4
PGUSER=$5

#------------------------------------------------
CONNECT_LDC="mssql -s $LDCSERVER -u $LDCUSER \
  -p $LDCPASS -d Footfall -e -f csv"
CONNECT_UCL="psql -q footfall -U $PGUSER"
VENDOR_AWK='{if(length($1)==8)
        {gsub(/:/,"",$1);
        print "\""$1"\",\""$2"\",\""$3"\""}}'
VENDOR_SQL='truncate vendors; \
  copy vendors from stdin with csv'
LOCATION_AWK='{addr=$2}
        {if(FNR>1){
        gsub(/"/,"",addr);
        split(addr,a," *, *");
        l=length(a);
        print $1","$2",\""a[l-2]"\",\""a[l]"\","$3",\
    "$4","$5","$6}}'
```

```
LOCATION_SQL='truncate locations; \
  copy locations from stdin with csv'
INSTALL_SQL='truncate installs; \
  copy installs from stdin with csv header'
DEVICE_SQL='truncate devices; \
  copy devices from stdin with csv header'
CALIB_SQL='truncate calibrations; \
  copy calibrations from stdin with csv header'

#------------------------------------------------
# Update all the tables
#------------------------------------------------
curl -s "$VENDORS" \
        | sed 's/"//g' \
        | awk -F $'\t' "$VENDOR_AWK" \
        | $CONNECT_UCL -c "$VENDOR_SQL"

$CONNECT_LDC -q "$(cat ./queries/locations)" \
        | awk -vFPAT='[^,]*|\"[^\"]*\"' \
    "$LOCATION_AWK" \
        | sed 's/""//g' \
        | $CONNECT_UCL -c "$LOCATION_SQL"

$CONNECT_LDC -q "$(cat ./queries/installs)" \
        | sed 's/""//g' \
        | $CONNECT_UCL -c "$INSTALL_SQL"

$CONNECT_LDC -q "$(cat ./queries/devices)" \
        | sed 's/""//g' \
        | $CONNECT_UCL -c "$DEVICE_SQL"

$CONNECT_LDC \
  -q "$(cat ./queries/calibrations)" \
        | sed 's/""//g' \
        | $CONNECT_UCL -c "$CALIB_SQL"
```

Rotate the salt value in the configuration file.

```
#! /bin/bash

```

```
#-----------------------------------------------
# Get the variable for the config file and
# generate a random string
#-----------------------------------------------
FILE=$1
SALT="\"$(cat /dev/urandom | tr -dc 'a-zA-Z0-9'\
  | fold -w 32 | head -n 1)\""
#-----------------------------------------------
# Change salt in the configuration file with the
# random string
#-----------------------------------------------
CONTENT=`jq ". + {salt:$SALT}" $FILE`
echo $CONTENT | jq '.' > $FILE
```

7.3.6 *SQL queries*

Manual counts.

```
select
        DeviceId as device,
        StartTime as start_time,
        EndTime as end_time,
        Count as count,
        Note as note
from calibrations;
```

Device information.

```
select
        Devices.Id as id,
        NfcLink as nfc,
        HardwareVersions.Name as hardware,
        SoftwareVersions.Version as software
from Devices
left join HardwareVersions on
   Devices.HardwareVersionId=HardwareVersions.Id
left join SoftwareVersions on
   Devices.SoftwareVersionId=SoftwareVersions.Id;
```

Location information.

```
select
        DeviceLocations.Id as id,
```

```
        Address as address,
        Lat as lat ,
        Lon as lon,
        DevicePositions.Name as position,
        DeviceLocationTypes.Name as type
from DeviceLocations
left join LdoPremises
        on DeviceLocations.PremiseId = LdoPremises.Id
left join DevicePositions
        on DeviceLocations.DevicePositionId = DevicePositions.Id
left join DeviceLocationTypes
        on DeviceLocations.DeviceLocationTypeId = DeviceLocationTypes.Id;
```

Installation notes.

```
select
        DeviceId as device,
        DeviceLocationId as location,
        FromDate as start_date,
        ToDate as end_date,
        Height as height,
        Depth as depth,
        Note as note
from DeviceHistories
left join DeviceInstallDetails on
        DeviceHistories.Id = DeviceInstallDetails.DeviceHistoryId
left join InstallSignatures on
        DeviceHistories.InstallSignatureId = InstallSignatures.Id
left join InstallNotes on
        InstallSignatures.Id = InstallNotes.Id;
```

7.4 *Benchmarking Data Tookit*

This section documents all the code that has been used in the research. The programming languages used are including but not limited to R, Bash, JavaScript and SQL.

7.4.1 *Simple implementation in R*

This R script lists all files in a given folder, parses them as JSON data serially, aggregates the records for each time interval and finally writes it to disk as a CSV file.

```
#! /usr/bin/Rscript
suppressMessages(library(tidyverse))
suppressMessages(library(RJSONIO))

day <- "~/unorganised-files/ff_sample/2018/01/01"
sensors <- paste(day, dir(day), sep = "/")[1:25]
probes <- NULL

for(sensor in sensors) {
  files <- paste(sensor, dir(sensor), sep = "/");
  for( file in files ) {
    records <- fromJSON(file);
    location <- vector(); signal <- vector();
    mac <- vector(); packets <- vector();
    vendor <- vector(); type <- vector(); time <- vector();
    for(record in records) {
      t <- strsplit(strsplit(file, '\\.')[[1]][1], '/')[[1]][8]
      l <- strsplit(strsplit(file, '\\.')[[1]][1], '/')[[1]][7]
      signal <- append(signal, record$Signal);
      mac <- append(mac, record$MacAddress);
      packets <- append(packets, record$PacketCount);
      type <- append(type, record$PacketType);
      vendor <- append(vendor, record$VendorMacPart);
      time <- append(time, t);
      location <- append(location, l);
    }
    recordsdf <- data.frame(location, time, signal,
                            mac, packets, type, vendor);
    if(is.null(probes)) { probes <- recordsdf; }
      else { probes <- rbind(probes, recordsdf); }
  }
}

probes %>%
  group_by(location, time) %>%
  summarise(count = length(unique(paste0(vendor, mac)))) %>%
  write.csv("output-old.csv", row.names = FALSE);
```

7.4.2 *Serial implementation in bash*

This bash script lists all the files in a given folder, parses them into JSON data *serially*, aggregates the resulting records for each time interval and finally writes it to disk as a CSV file.

```
#! /bin/bash
awkc="awk -vFPAT='[^,]*|\"[^\"]*\"' -v OFS=','"
FOLDER="/home/ucfnbso/unorganised-files/ff_sample/2018/01/01/"
SENSORS=`ls $FOLDER | head -n $1`

for SENSOR in $SENSORS;
do
  jq_string=".[] | \
    [\"$SENSOR\",\
    ( (input_filename/\"/\" | .[ .|length-1 ] )/\".\" | .[0]),\
    .VendorMacPart+.MacAddress] \
    | @csv";
  cmd="jq -r '$jq_string' $FOLDER$SENSOR/*.pd \
    | sort | uniq \
    | $awkc '{print \$1,\$2}' \
    | sort | uniq -c";
  echo "$(eval $cmd)" > output-new.csv;
done
```

7.4.3 *Parallel implementation in bash*

This bash script lists all the files in a given folder, parses them into JSON data *in parallel*, aggregates the resulting records for each time interval and finally writes it to disk as a CSV file.

```
#! /bin/bash
awkc="awk -vFPAT='[^,]*|\"[^\"]*\"' -v OFS=','"
folder="/home/ucfnbso/unorganised-files/ff_sample/2018/01/01/"
sensors=`ls $folder | head -n $1`

jq_string=".[] | \
  [\"{}\",\
  ( (input_filename/\"/\" | .[ .|length-1 ] )/\".\" | .[0]),\
  .VendorMacPart+.MacAddress] \
  | @csv";
```

```
cmd="jq -r '$jq_string' $folder{}/*.pd \
  | sort | uniq \
  | $awkc '{print \$1,\$2}' \
  | sort | uniq -c";

echo "$sensors" \
  | parallel "$cmd" \
  > output-new-parallel.csv
```

7.5 *Sample Probe Request*

This is a sample probe request captured using tshark and saved in the JSON format.


```
{
   "_index": "packets-2017-11-15",
   "_type": "pcap_file",
   "_score": null,
   "_source": {
     "layers": {
       "frame": {
         "frame.interface_id": "0",
         "frame.interface_id_tree": {
           "frame.interface_name": "en0"
         },
         "frame.encap_type": "23",
         "frame.time": "Nov 15, 2017 18:46:56.155602000 GMT",
         "frame.offset_shift": "0.000000000",
         "frame.time_epoch": "1510771616.155602000",
         "frame.time_delta": "0.019159000",
         "frame.time_delta_displayed": "0.019159000",
         "frame.time_relative": "0.343422000",
         "frame.number": "6",
         "frame.len": "142",
         "frame.cap_len": "142",
         "frame.marked": "0",
         "frame.ignored": "0",
         "frame.protocols": "radiotap:wlan_radio:wlan"
       },
       "radiotap": {
         "radiotap.version": "0",
         "radiotap.pad": "0",
         "radiotap.length": "25",
         "radiotap.present": {
           "radiotap.present.word": "0x0000086f",
           "radiotap.present.word_tree": {
             "radiotap.present.tsft": "1",
             "radiotap.present.flags": "1",
```

```
            "radiotap.present.rate": "1",
            "radiotap.present.channel": "1",
            "radiotap.present.fhss": "0",
            "radiotap.present.dbm_antsignal": "1",
            "radiotap.present.dbm_antnoise": "1",
            "radiotap.present.lock_quality": "0",
            "radiotap.present.tx_attenuation": "0",
            "radiotap.present.db_tx_attenuation": "0",
            "radiotap.present.dbm_tx_power": "0",
            "radiotap.present.antenna": "1",
            "radiotap.present.db_antsignal": "0",
            "radiotap.present.db_antnoise": "0",
            "radiotap.present.rxflags": "0",
            "radiotap.present.xchannel": "0",
            "radiotap.present.mcs": "0",
            "radiotap.present.ampdu": "0",
            "radiotap.present.vht": "0",
            "radiotap.present.timestamp": "0",
            "radiotap.present.reserved": "0x00000000",
            "radiotap.present.rtap_ns": "0",
            "radiotap.present.vendor_ns": "0",
            "radiotap.present.ext": "0"
          }
        },
        "radiotap.mactime": "836459236",
        "radiotap.flags": "0x00000012",
        "radiotap.flags_tree": {
          "radiotap.flags.cfp": "0",
          "radiotap.flags.preamble": "1",
          "radiotap.flags.wep": "0",
          "radiotap.flags.frag": "0",
          "radiotap.flags.fcs": "1",
          "radiotap.flags.datapad": "0",
          "radiotap.flags.badfcs": "0",
          "radiotap.flags.shortgi": "0"
        },
        "radiotap.datarate": "6",
        "radiotap.channel.freq": "5180",
        "radiotap.channel.flags": "0x00000140",
```

```
          "radiotap.channel.flags_tree": {
            "radiotap.channel.flags.turbo": "0",
            "radiotap.channel.flags.cck": "0",
            "radiotap.channel.flags.ofdm": "1",
            "radiotap.channel.flags.2ghz": "0",
            "radiotap.channel.flags.5ghz": "1",
            "radiotap.channel.flags.passive": "0",
            "radiotap.channel.flags.dynamic": "0",
            "radiotap.channel.flags.gfsk": "0",
            "radiotap.channel.flags.gsm": "0",
            "radiotap.channel.flags.sturbo": "0",
            "radiotap.channel.flags.half": "0",
            "radiotap.channel.flags.quarter": "0"
          },
          "radiotap.dbm_antsignal": "-76",
          "radiotap.dbm_antnoise": "-96",
          "radiotap.antenna": "1"
        },
        "wlan_radio": {
          "wlan_radio.phy": "5",
          "wlan_radio.11a.turbo_type": "0",
          "wlan_radio.data_rate": "6",
          "wlan_radio.channel": "36",
          "wlan_radio.frequency": "5180",
          "wlan_radio.signal_dbm": "-76",
          "wlan_radio.noise_dbm": "-96",
          "wlan_radio.timestamp": "836459236",
          "wlan_radio.duration": "180",
          "wlan_radio.duration_tree": {
            "wlan_radio.preamble": "20",
            "wlan_radio.ifs": "19096",
            "wlan_radio.start_tsf": "836459056",
            "wlan_radio.end_tsf": "836459236"
          }
        },
        "wlan": {
          "wlan.fc.type_subtype": "4",
          "wlan.fc": "0x00004000",
          "wlan.fc_tree": {
```

```
            "wlan.fc.version": "0",
            "wlan.fc.type": "0",
            "wlan.fc.subtype": "4",
            "wlan.flags": "0x00000000",
            "wlan.flags_tree": {
              "wlan.fc.ds": "0x00000000",
              "wlan.fc.tods": "0",
              "wlan.fc.fromds": "0",
              "wlan.fc.frag": "0",
              "wlan.fc.retry": "0",
              "wlan.fc.pwrmgt": "0",
              "wlan.fc.moredata": "0",
              "wlan.fc.protected": "0",
              "wlan.fc.order": "0"
            }
          },
          "wlan.duration": "0",
          "wlan.ra": "ff:ff:ff:ff:ff:ff",
          "wlan.ra_resolved": "Broadcast",
          "wlan.da": "ff:ff:ff:ff:ff:ff",
          "wlan.da_resolved": "Broadcast",
          "wlan.ta": "94:b1:0a:79:15:9b",
          "wlan.ta_resolved": "SamsungE_79:15:9b",
          "wlan.sa": "94:b1:0a:79:15:9b",
          "wlan.sa_resolved": "SamsungE_79:15:9b",
          "wlan.bssid": "ff:ff:ff:ff:ff:ff",
          "wlan.bssid_resolved": "Broadcast",
          "wlan.addr": "ff:ff:ff:ff:ff:ff",
          "wlan.addr_resolved": "Broadcast",
          "wlan.addr": "94:b1:0a:79:15:9b",
          "wlan.addr_resolved": "SamsungE_79:15:9b",
          "wlan.addr": "ff:ff:ff:ff:ff:ff",
          "wlan.addr_resolved": "Broadcast",
          "wlan.frag": "0",
          "wlan.seq": "2533",
          "wlan.fcs": "0x5a6629b7",
          "wlan.fcs.status": "1"
        },
        "wlan": {
```

```
            "wlan.tagged.all": {
              "wlan.tag": {
                "wlan.tag.number": "0",
                "wlan.tag.length": "9",
                "wlan.ssid": "VM3280449"
              },
              "wlan.tag": {
                "wlan.tag.number": "1",
                "wlan.tag.length": "8",
                "wlan.supported_rates": "12",
                "wlan.supported_rates": "18",
                "wlan.supported_rates": "24",
                "wlan.supported_rates": "36",
                "wlan.supported_rates": "48",
                "wlan.supported_rates": "72",
                "wlan.supported_rates": "96",
                "wlan.supported_rates": "108"
              },
              "wlan.tag": {
                "wlan.tag.number": "45",
                "wlan.tag.length": "26",
                "wlan.ht.capabilities": "0x00000062",
                "wlan.ht.capabilities_tree": {
                  "wlan.ht.capabilities.ldpccoding": "0",
                  "wlan.ht.capabilities.width": "1",
                  "wlan.ht.capabilities.sm": "0x00000000",
                  "wlan.ht.capabilities.green": "0",
                  "wlan.ht.capabilities.short20": "1",
                  "wlan.ht.capabilities.short40": "1",
                  "wlan.ht.capabilities.txstbc": "0",
                  "wlan.ht.capabilities.rxstbc": "0x00000000",
                  "wlan.ht.capabilities.delayedblockack": "0",
                  "wlan.ht.capabilities.amsdu": "0",
                  "wlan.ht.capabilities.dsscck": "0",
                  "wlan.ht.capabilities.psmp": "0",
                  "wlan.ht.capabilities.40mhzintolerant": "0",
                  "wlan.ht.capabilities.lsig": "0"
                },
                "wlan.ht.ampduparam": "0x00000017",
```

```
            "wlan.ht.ampduparam_tree": {
              "wlan.ht.ampduparam.maxlength": "0x00000003",
              "wlan.ht.ampduparam.mpdudensity": "0x00000005",
              "wlan.ht.ampduparam.reserved": "0x00000000"
            },
            "wlan.ht.mcsset": "MCS Set",
            "wlan.ht.mcsset_tree": {
              "wlan.ht.mcsset.rxbitmask": {
                "wlan.ht.mcsset.rxbitmask.0to7": "0x000000ff",
                "wlan.ht.mcsset.rxbitmask.8to15": "0x00000000",
                "wlan.ht.mcsset.rxbitmask.16to23": "0x00000000",
                "wlan.ht.mcsset.rxbitmask.24to31": "0x00000000",
                "wlan.ht.mcsset.rxbitmask.32": "0x00000001",
                "wlan.ht.mcsset.rxbitmask.33to38": "0x00000000",
                "wlan.ht.mcsset.rxbitmask.39to52": "0x00000000",
                "wlan.ht.mcsset.rxbitmask.53to76": "0x00000000"
              },
              "wlan.ht.mcsset.highestdatarate": "0x00000000",
              "wlan.ht.mcsset.txsetdefined": "0",
              "wlan.ht.mcsset.txrxmcsnotequal": "0",
              "wlan.ht.mcsset.txmaxss": "0x00000000",
              "wlan.ht.mcsset.txunequalmod": "0"
            },
            "wlan.htex.capabilities": "0x00000000",
            "wlan.htex.capabilities_tree": {
              "wlan.htex.capabilities.pco": "0",
              "wlan.htex.capabilities.transtime": "0x00000000",
              "wlan.htex.capabilities.mcs": "0x00000000",
              "wlan.htex.capabilities.htc": "0",
              "wlan.htex.capabilities.rdresponder": "0"
            },
            "wlan.txbf": "0x00000000",
            "wlan.txbf_tree": {
              "wlan.txbf.txbf": "0",
              "wlan.txbf.rxss": "0",
              "wlan.txbf.txss": "0",
              "wlan.txbf.rxndp": "0",
              "wlan.txbf.txndp": "0",
              "wlan.txbf.impltxbf": "0",
```

```
                "wlan.txbf.calibration": "0x00000000",
                "wlan.txbf.csi": "0",
                "wlan.txbf.fm.uncompressed.tbf": "0",
                "wlan.txbf.fm.compressed.tbf": "0",
                "wlan.txbf.rcsi": "0x00000000",
                "wlan.txbf.fm.uncompressed.rbf": "0x00000000",
                "wlan.txbf.fm.compressed.bf": "0x00000000",
                "wlan.txbf.mingroup": "0x00000000",
                "wlan.txbf.csinumant": "0x00000000",
                "wlan.txbf.fm.uncompressed.maxant": "0x00000000",
                "wlan.txbf.fm.compressed.maxant": "0x00000000",
                "wlan.txbf.csi.maxrows": "0x00000000",
                "wlan.txbf.channelest": "0x00000000",
                "wlan.txbf.reserved": "0x00000000"
              },
              "wlan.asel": "0x00000000",
              "wlan.asel_tree": {
                "wlan.asel.capable": "0",
                "wlan.asel.txcsi": "0",
                "wlan.asel.txif": "0",
                "wlan.asel.csi": "0",
                "wlan.asel.if": "0",
                "wlan.asel.rx": "0",
                "wlan.asel.sppdu": "0",
                "wlan.asel.reserved": "0x00000000"
              }
            },
            "wlan.tag": {
              "wlan.tag.number": "107",
              "wlan.tag.length": "1",
              "wlan.interworking.access_network_type": "15",
              "wlan.interworking.internet": "0",
              "wlan.interworking.asra": "0",
              "wlan.interworking.esr": "0",
              "wlan.interworking.uesa": "0"
            },
            "wlan.tag": {
              "wlan.tag.number": "221",
              "wlan.tag.length": "5",
```

```
                "wlan.tag.oui": "5271450",
                "wlan.tag.vendor.oui.type": "16",
                "wlan.hs20.indication.dgaf_disabled": "0",
                "wlan.hs20.indication.pps_mo_id_present": "0",
                "wlan.hs20.indication.anqp_domain_id_present": "0",
                "wlan.hs20.indication.release_number": "1"
              },
              "wlan.tag": {
                "wlan.tag.number": "221",
                "wlan.tag.length": "7",
                "wlan.tag.oui": "20722",
                "wlan.tag.vendor.oui.type": "8",
                "wlan.wfa.ie.type": "0x00000008"
              },
              "wlan.tag": {
                "wlan.tag.number": "221",
                "wlan.tag.length": "9",
                "wlan.tag.oui": "4120",
                "wlan.tag.vendor.oui.type": "2",
                "wlan.tag.vendor.data": "02:00:00:10:00:00"
              }
} } } } }
```

7.6 Open-source Software Used

This section provides a non-exhaustive list of the key open-source/free software that have been used in this research.

- **R** - programming language for statistical computing
  - **tidyverse** - An opinionated collection of R packages designed for data science.
  - **imputeTS** - Package for imputation missing values in univariate time series.
  - **tmap** - A flexible, layer-based, and easy to use package to create thematic maps.
  - **lubridate** - Package for working with date-times and time-spans
  - **ggplot2** - A system for declaratively creating graphics, based on The Grammar of Graphics.
  - **classInt** - Package for choosing univariate class intervals for mapping or other graphics purposes.
  - **Cairo** - 2D graphics library with support for multiple output devices.
  - **fmsb** - Package with methods and functions for demographic analysis.
  - **digest** - Package for the creation of hash digests of arbitrary R objects.
  - **ggrepel** - Package that provides geoms for ggplot2 to repel overlapping text labels.
  - **ggridges** - Package for creating ridge plots.
  - **maptools** - Package for manipulating geographic data.
  - **tidyquant** - Package that brings financial analysis and charting to tidyverse.
  - **treemapify** - Package for creating tree-maps.
  - **spatial features** - Package for manipulating geographic data within tidyverse.
  - **RJSONIO** - Package for manipulating JSON objects.
  - **rgdal** - Wrapper for the Geospatial Data Abstraction Library.
  - **rgeos** - Wrapper for the Geometry Engine - Open Source.
  - **viridis** - Package providing pretty color scales for visualisations.

  - **xtable** - Package for coercing data to LaTeX and HTML tables.
  - **scales** - Package for providing graphical scales mapping data to aesthetics.
  - **showtext** - Package for managing fonts.
  - **reshape2** - Package for transforming data between long and wide format.
  - **rmarkdown** - Package for integrating markdown with R assisting reproducible research.

- **Python** - An interpreted, high-level, general-purpose programming language.
- **PHP** - A general-purpose programming language originally designed for server-side web development.
- **JavaScript** - A high-level, interpreted scripting language for client-side web development.
  - **node.js** - JavaScript based runtime built on chrome's V8 engine.
  - **socket.io** - web sockets implementation for real-time, bidirectional and event-based communication.
  - **moment.js** - JavaScript library for dealing with date-time and time-spans.
  - **pm2** - Process management library for working with node.js applications.
  - **express** - A fast, unopinionated, minimalist web framework for Node.js
  - **Data Driven Documents** - A JavaScript library for visualizing data with HTML, SVG, and CSS.
  - **jQuery** - A JavaScript library designed to simplify HTML DOM manipulation.
  - **Bootstrap** - A framework for building responsive, mobile first websites.
  - **highcharts** - A JavaScript library for drawing interactive charts from data.
- **GNU/Linux** - An operating system and an extensive collection of open source and free computer software.
  - **Arch Linux** - A lightweight and flexible Linux distribution.

  - **CentOS** - A community supported computing platform compatible with Red Hat Enterprise Linux.
  - **Debian** - A Linux distribution focussing on stability.
  - **Ubuntu** - A Debian based Linux distribution focussing on ease of use.
  - **Alpine Linux** - Ultra minimalistic Linux distribution focussing on resource efficiency.
- **git** - A simple distributed version control system.
- **imagemagik** - Suite for displaying, converting and editing images.
- **ffmpeg** - Suite for converting and editing video files.
- **fzf** - A general-purpose command-line fuzzy finder.
- **ripgrep** - Rust based grep implementation for searching the content of files.
- **MySQL** - A relational database management system focussing on speed and ease of use.
- **Postgres** - A relational database management system emphasizing extensibility and technical standards compliance.
- **PostGIS** - Extension providing spatial objects for the PostgreSQL database.
- **QGIS** - Geographic Information System for creating, editing, visualising, analysing and publishing geospatial information.
- **gdal** - Geographic Data abstraction library.
- **geos** - Geometry Engine Open Source.
- **igraph** - R and Python library for dealing with networks / Graphs.
- **OpenStreetMap** - A collaborative project to create a free editable map of the world.
- **Leaflet** - A JavaScript library for mobile-friendly interactive maps.
- **Android** - Open source mobile operating system based on Linux.
- **vim** - Vim is a highly configurable, modal text editor built with focus on efficient.
- **Latex** - A high quality professional typesetting system.

- **jq** - A command line based JSON processor.
- **Apache** - A feature rich web server.
- **nginx** - A asynchronous, event-driven web server focussing on resource efficiency.
- **OpenJDK** - An open source implementation of the Java Platform.
- **Wireshark** - A free and open-source packet analyzer
- **OpenSSH** - A connectivity tool for remote login with the SSH protocol.
- **OpenSSL** - A full-featured toolkit for the Transport Layer Security and Secure Sockets Layer protocols.
- **GNUPG** - A complete and free implementation of the OpenPGP standard.
- **gnu-parallel** - A shell tool for executing jobs in parallel using one or more computers.
- **Libreoffice** - A free and open-source office suite built by The Document Foundation.
- **RaspberryPi** - A series of low-cost, flexible single-board computers.
- **Docker** - A platform for doing OS level virtualisation for delivering software.
- **termux** - A terminal emulator and Linux environment for Android.

Bibliography

Michael Abbott-Jard, Harpal Shah, and Ashish Bhaskar. Empirical evaluation of bluetooth and wifi scanning for road transport. In *Australasian Transport Research Forum (ATRF), 36th, 2013, Brisbane, Queensland, Australia*, page 14, 2013.

Naeim Abedi, Ashish Bhaskar, and Edward Chung. Bluetooth and wi-fi mac address based crowd data collection and monitoring : Benefits , challenges and enhancement. In *Australasian Transport Research Forum 2013 Proceedings 2*, pages 1–17, 2013. URL http://www.patrec.org/atrf.aspx.

John M Abowd, John Haltiwanger, and Julia Lane. Integrated longitudinal employer-employee data for the united states. *The American Economic Review*, 94(2):224–229, 2004. ISSN 0002-8282.

Florian Adamsky, Tatiana Retunskaia, Stefan Schiffner, Christian Köbel, and Thomas Engel. Wlan device fingerprinting using channel state information (csi). In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 277–278. ACM, 2018.

Rein Ahas and Ülar Mark. Location based services - new challenges for planning and public administration? *Futures*, 37(6):547–561, 2005. ISSN 00163287. DOI: 10.1016/j.futures.2004.10.012.

Mohammed Alahmadi, Peter Atkinson, and David Martin. Estimating the spatial distribution of the population of riyadh, saudi arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41:167–176, 2013. ISSN 0198-9715.

David E Alexander. *Principles of emergency planning and management*. Oxford University Press on Demand, 2002. ISBN 0195218388.

Alexander Amini, Kevin Kung, Chaogui Kang, Stanislav Sobolevsky, and Carlo Ratti. The impact of social segregation on human mobility

in developing and industrialized regions. *EPJ Data Science*, 3(1):1, 2014. ISSN 2193-1127.

Aleksandar Antonic, Ivana Podnar Zarko, and Domagoj Jakobovic. Inferring presence status on smartphones: The big data perspective. In *Proceedings - International Symposium on Computers and Communications*, pages 600–605, 2013. ISBN 9781479937554. DOI: 10.1109/ISCC.2013.6755013.

Irina Arhipova, Gundars Berzins, Edgars Brekis, Martins Opmanis, Juris Binde, Jevgenija Kravcova, and Inna Steinbuka. Pattern identification by factor analysis for regions with similar economic activity based on mobile communication data. In *Future of Information and Communication Conference*, pages 561–569. Springer, 2018.

Kai O Arras, Slawomir Grzonka, Matthias Luber, and Wolfram Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 1710–1715. IEEE, 2008. ISBN 1424416469.

Dani Arribas-Bel and Emmanouil Tranos. New approaches to measure the spatial structure (s) of cities. In *Proceedings of GISRUK 2015*. GIS Research UK (GISRUK), 2015.

Daniel Arribas-Bel. Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49:45–53, 2014. ISSN 01436228. DOI: 10.1016/j.apgeog.2013.09.012.

Daniel Arribas-Bel and Fernando Sanz-Gracia. The validity of the monocentric city model in a polycentric age: Us metropolitan areas in 1990, 2000 and 2010. *Urban Geography*, 35(7):980–997, 2014. ISSN 0272-3638.

Yasuo Asakura and Eiji Hato. Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies*, 12(3):273–291, 2004. ISSN 0968-090X.

Fereshteh Asgari, Vincent Gauthier, and Monique Becker. A survey on human mobility and its applications. *arXiv preprint arXiv:1307.0814*, 2013.

Paramvir Bahl and Venkata N Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *INFOCOM 2000. Nineteenth*

Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, volume 2, pages 775–784. Ieee, 2000. ISBN 0780358805.

Udana Bandara, Mikio Hasegawa, Masugi Inoue, Hiroyuki Morikawa, and Tomonori Aoyama. Design and implementation of a bluetooth signal strength based location sensing system. In *Radio and Wireless Conference, 2004 IEEE*, pages 319–322. IEEE, 2004. ISBN 0780384512.

Tengfei Bao, Huanhuan Cao, Qiang Yang, Enhong Chen, and Jilei Tian. Mining significant places from cell id trajectories: A geo-grid based approach. In *2012 IEEE 13th International Conference on Mobile Data Management*, pages 288–293. IEEE, 2012. ISBN 1467317969.

Marco V Barbera, Alessandro Epasto, Alessandro Mei, Vasile C Perta, and Julinda Stefa. Signals from the crowd: uncovering social relationships through smartphone probes. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 265–276. ACM, 2013. ISBN 145031953X.

Michael Batty. Invisible cities. *Environment and Planning B: Planning and Design*, 17(2):127–130, 1990.

Michael Batty. Virtual geography. *Futures*, 29(4-5):337–352, 1997. ISSN 00163287. DOI: 10.1016/S0016-3287(97)00018-9. URL http://www.sciencedirect.com/science/article/pii/S0016328797000189.

Michael Batty. The pulse of the city. *Environment and Planning B: Planning and Design*, 37(4):575–577, 2010. ISSN 0265-8135.

Michael Batty. *The new science of cities*. Mit Press, 2013a. ISBN 0262019523.

Michael Batty. The future cities agenda. *Environ. Plann. B Plann. Des*, 40:191–194, 2013b.

Michael Batty, Kay W Axhausen, Fosca Giannotti, Alexei Pozdnoukhov, Armando Bazzani, Monica Wachowicz, Georgios Ouzounis, and Yuval Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012. ISSN 1951-6355.

Anil Bawa-Cavia. Sensing the urban: using location-based social network data in urban analysis. In *Workshop on Pervasive and Urban Applications (PURBA)*, pages 1–7, 2011.

Johannes K Becker, David Li, and David Starobinski. Tracking anonymized bluetooth devices. *Proceedings on Privacy Enhancing Technologies*, 1:17, 2019.

Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Route classification using cellular handoff patterns. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 123–132. ACM, 2011a. ISBN 1450306306.

Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011b. ISSN 1536-1268. DOI: 10.1109/MPRV.2011.44. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5928310.

Alan Bensky. Wireless positioning technologies and applications, artech house. *Inc., Norwood, MA*, 2007.

Luca Bertolini and Martin Dijst. Mobility environments and network cities. *Journal of urban design*, 8(1):27–43, 2003.

Luís M A Bettencourt. The origins of scaling in cities. *science*, 340(6139): 1438–1441, 2013. ISSN 0036-8075.

Luis M A Luís M.A. Bettencourt. The uses of big data in cities. *Big Data*, 2(1):12–22, 2014. ISSN 2167-6461. DOI: 10.1089/big.2013.0042. URL http://online.liebertpub.com/doi/abs/10.1089/big.2013.0042.

Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Jerome Dobson. Landscan. *Geoinformatics*, 5(2):34–37, 2002.

Budhendra Bhaduri, Eddie Bright, and Phil Coleman. Development of a high resolution population dynamics model. *Geocomputation 2005*, 2005.

Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Marie L. Urban. Landscan usa: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. In *GeoJournal*, volume 69, pages 103–117. Springer, 2007. ISBN 0165-0009. DOI: 10.1007/s10708-007-9105-9.

Bastian Bloessl, Christoph Sommer, Falko Dressier, and David Eckhoff. The scrambler attack: A robust physical layer attack on location privacy

in vehicular networks. In *Computing, Networking and Communications (ICNC), 2015 International Conference on*, pages 395–400. IEEE, 2015. ISBN 1479969591.

R Bolla, F Davoli, and F Giordano. Estimating road traffic parameters from mobile communications. In *Proceedings 7th World Congress on ITS, Turin, Italy*, 2000.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. URL http://vis.stanford.edu/papers/d3.

Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006. ISSN 0028-0836.

Caroline O Buckee, Andrew J Tatem, Justin Lessler, Ottar N Bjornstad, Bryan T Grenfell, Janeth Kombich, Nathan Eagle, C J E Metcalf, and Amy Wesolowski. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences*, 112(35):11114–9, 2015. ISSN 0027-8424. DOI: 10.1073/pnas.1423542112. URL http://doi.org/10.1073/pnas.1423542112.

Nirupama Bulusu, John Heidemann, and Deborah Estrin. Gps-less low-cost outdoor localization for very small devices. *IEEE personal communications*, 7(5):28–34, 2000. ISSN 1070-9916.

N Caceres, J P Wideberg, and F G Benitez. Deriving origin destination data from a mobile phone network. *IET Intelligent Transport Systems*, 1 (1):15–26, 2007. ISSN 1751-956X.

Quin Cai and Jake K Aggarwal. Tracking human motion using multiple cameras. In *International Conference on Pattern Recognition*, volume 13, pages 68–72. Citeseer, 1996. ISBN 1051-4651.

Francesco Calabrese, Francisco C Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *International Conference on Pervasive Computing*, pages 22–37. Springer, 2010.

Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44, 2011. ISSN 15361268. DOI: 10.1109/MPRV.2011.41.

Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Jr. Ferreira Joseph, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26(0):301–313, 2013.

Francesco Calabrese, Laura Ferrari, and Vincent D. Blondel. Urban sensing using mobile phone network data: a survey of research. *ACM Computing Surveys (CSUR)*, 47(2):25, 2015. ISSN 0360-0300. DOI: 10.1145/2655691. URL http://dl.acm.org/citation.cfm?doid=2658850.2655691.

Andrew T Campbell, Shane B Eisenman, Kristíf Fodor, Nicholas D Lane, Hong Lu, Emiliano Miluzzo, Mirco Musolesi, Ronald A Peterson, and Xiao Zheng. Transforming the social networking experience with sensing presence from mobile phones. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 367–368. ACM, 2008.

Julián Candia, Marta C. González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A: Math. Theor*, 41(22):224015–11, 2008. ISSN 1751-8113. DOI: 10.1088/1751-8113/41/22/224015.

M Castells. Grassrooting the space of flows. i wheeler, aoyama and warf [eds.] cities in the telecommunications age, 2000.

M Castells. *The Rise of the Network Society*, volume I. Massachusetts: Blackwell Publishing, 2010. ISBN 9781405196864. DOI: 10.2307/1252090. URL http://www.lavoisier.fr/livre/notice.asp?depuis=e.lavoisier.fr{&}id=9781405196864.

Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S Ebert, and Thomas Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 143–152. IEEE, 2012. ISBN 1467347523.

Jie Chen, Tao Pei, Shih-Lung Shaw, Feng Lu, Mingxiao Li, Shifen Cheng, Xiliang Liu, and Hengcai Zhang. Fine-grained prediction of urban population using mobile phone location data. *International Journal of Geographical Information Science*, 32(9):1770–1786, 2018.

K Chen. An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing*, 23(1):37–48, 2002. ISSN 0143-1161.

Linsong Cheng and Jiliang Wang. How can i guard my ap?: non-intrusive user identification for mobile devices using wifi signals. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 91–100. ACM, 2016.

Ningning Cheng, Prasant Mohapatra, Mathieu Cunche, Mohamed Ali Kaafar, Roksana Boreli, and Srikanth Krishnamurthy. Inferring user relationship from hidden information in wlans. In *Military Communications Conference, 2012-MILCOM 2012*, pages 1–6. IEEE, 2012. ISBN 1467317314.

Peng Cheng, Zhijun Qiu, and Bin Ran. Particle filter based traffic state estimation using cell phone network data. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 1047–1052. IEEE, 2006. ISBN 1424400937.

Guanghua Chi, Jean-Claude Thill, Daoqin Tong, Li Shi, and Yu Liu. Uncovering regional characteristics from mobile phone data: A network science approach. *Papers in Regional Science*, 2014. ISSN 1435-5957.

Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

Tom Chothia and Vitaliy Smirnov. A traceability attack against e-passports. In *International Conference on Financial Cryptography and Data Security*, pages 20–34. Springer, 2010. ISBN 3642145760.

Cisco. *Wi-Fi Location-Based Services 4.1 Design Guide*. Cisco Systems Inc., 2008. URL https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Mobility/WiFiLBS-DG.html.

Gerald Combs and Contributors. Wireshark - network protocol analyzer. https://www.wireshark.org/about.html, 2018.

Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

Tomas Crols and Nick Malleson. Quantifying the ambient population using hourly population footfall data and an agent-based model of daily mobility. *Geoinformatica*, pages 1–20, 2019.

Mathieu Cunche. I know your mac address: targeted tracking of individual using wi-fi, 2014. ISSN 22638733.

Mathieu Cunche, Mohamed Ali Kaafar, and Roksana Boreli. I know who you will meet this evening! linking wireless devices using wi-fi probe requests. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a*, pages 1–9. IEEE, 2012. ISBN 1467312398.

Mathieu Cunche, Mohamed-Ali Kaafar, and Roksana Boreli. Linking wireless devices using information contained in wi-fi probe requests. *Pervasive and Mobile Computing*, 11:56–69, 2014. ISSN 1574-1192.

S Cutter, Douglas B Richardson, and Thomas J Wilbanks. The changing landscape of fear. *The geographical dimensions of terrorism*, pages 1–5, 2006.

Peng Dai, Yuan Yang, Manyi Wang, and Ruqiang Yan. Combination of dnn and improved knn for indoor location fingerprinting. *Wireless Communications and Mobile Computing*, 2019, 2019.

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

Deloitte. Mobile consumer survey - united kingdom, 2018. URL `https://www.deloitte.co.uk/mobileuk/`.

Merkebe Getachew Demissie, Gonçalo Homem de Almeida Correia, and Carlos Bento. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. *Transportation research part C: emerging technologies*, 32:76–88, 2013. ISSN 0968-090X.

Zhihong Deng, Yun Cao, Pengyu Wang, and Bo Wang. An improved heuristic drift elimination method for indoor pedestrian positioning. *Sensors*, 18(6):1874, 2018.

Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J

Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014. ISSN 0027-8424. DOI: 10.1073/pnas.1408439111.

Adriano Di Luzio, Alessandro Mei, and Julinda Stefa. Mind your probes: De-anonymization of large crowds through smartphone wifi probe requests. In *Proceedings - IEEE INFOCOM*, volume 2016-July, 2016. ISBN 9781467399531. DOI: 10.1109/INFOCOM.2016.7524459.

Bram Dil and Paul J M Havinga. Stochastic radio interferometric positioning in the 2.4 ghz range. In *SenSys*, pages 108–120, 2011.

Vaibhav Dinesh, Prakash Tripathi, and Ajit Singh. Device-location estimation based on rssi measurements over wifi and bluetooth, March 2 2017. US Patent App. 15/056,902.

Trinh Minh Tri Do and Daniel Gatica-Perez. The places of our lives: Visiting patterns and automatic labeling from longitudinal smartphone data. *IEEE Transactions on Mobile Computing*, 13(3):638–648, 2014. ISSN 15361233. DOI: 10.1109/TMC.2013.19.

Jerome E Dobson and Peter F Fisher. Geoslavery. *IEEE Technology and Society Magazine*, 22(1):47–52, 2003. ISSN 0278-0097.

Jerome E Dobson, Edward A Bright, Phillip R Coleman, Richard C Durfee, and Brian A Worley. Landscan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing*, 66(7):849–857, 2000. ISSN 0099-1112.

Jerome E Dobson, Eddie A Bright, Phil R Coleman, and Budhendra L Bhaduri. Landscan2000: A new global population geography. Technical report, Oak Ridge National Lab.(ORNL), 2003.

Adam Drake. Command-line tools can be 235x faster than your hadoop cluster, Jan 2014. URL `https://bit.ly/2s2XZYI`.

Matt Duckham and Lars Kulik. Location privacy and location-aware computing. *Dynamic & mobile GIS: investigating change in space and time*, 3:35–51, 2006.

Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009. ISSN 03405443. DOI: 10.1007/s00265-009-0739-0.

Nature Editorial. A flood of hard data. *Nature*, 435:698, 2008. ISSN 0028-0836. DOI: 10.1038/453698a.

A Elgohary. On detecting device-free entities using wifi signals. *ece.uwaterloo.ca*, 2013. URL https://ece.uwaterloo.ca/{~}aelgohar/stat841-report.pdf{%}5Cnpapers3://publication/uuid/D6821814-0041-47E6-9A26-96A32F41B07F.

Gorkem Erinc, Benjamin Balaguer, and Stefano Carpin. Heterogeneous map merging using wifi signals. In *IEEE International Conference on Intelligent Robots and Systems*, pages 5258–5264, 2013. ISBN 9781467363587. DOI: 10.1109/IROS.2013.6697117.

Sara Irina Fabrikant. Towards an understanding of geovisualization with dynamic displays: Issues and prospects. In *AAAI Spring Symposium: Reasoning with Mental and External Diagrams: Computational Modeling and Spatial Assistance*, pages 6–11, 2005.

Katayoun Farrahi and Daniel Gatica-Perez. Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing*, 4(4):746–755, 2010.

Emilio Ferrara, Pasquale De Meo, Salvatore Catanese, and Giacomo Fiumara. Visualizing criminal networks reconstructed from mobile phone records. In *CEUR Workshop Proceedings*, volume 1210, 2014.

Donald L Foley. Urban daytime population: a field for demographic-ecological analysis. *Social Forces*, pages 323–330, 1954. ISSN 0037-7732.

Jason Franklin, Damon McCoy, Parisa Tabriz, Vicentiu Neagoe, Jamie V Randwyk, and Douglas Sicker. Passive data link layer 802.11 wireless device driver fingerprinting. In *USENIX Security Symposium*, volume 3, pages 16–89, 2006.

Julien Freudiger. How talkative is your mobile device?: An experimental study of Wi-Fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '15, pages 8:1–8:6, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3623-9. DOI: 10.1145/2766498.2766517. URL http://doi.acm.org/10.1145/2766498.2766517.

Yuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. A pedestrian flow analysis system using wi-fi packet sensors to a real environment. *Proceedings of UbiComp-Adjunct*, pages 721–730, 2014. DOI: 10.1145/2638728.2641312. URL http://ubicomp.org/ubicomp2014/proceedings/ubicomp{_}adjunct/workshops/HASCA/p721-fukuzaki.pdf.

Yuuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. Statistical analysis of actual number of pedestrians for wi-fi packet-based pedestrian flow sensing. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 1519–1526, 2015. ISBN 978-1-4503-3575-1. DOI: 10.1145/2800835.2801623. URL http://doi.acm.org/10.1145/2800835.2801623.

Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015.

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, pages 20–43, Bolton Landing, NY, 2003.

Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 330–339. ACM, 2007. ISBN 1595936092.

Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal—The International Journal on Very Large Data Bases*, 20(5):695–719, 2011. ISSN 1066-8888.

Fabien Girardin, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing*, 7(4):36–43, 2008. ISSN 1536-1268.

Fabien Girardin, Andrea Vaccari, Alexandre Gerber, Assaf Biderman, and Carlo Ratti. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.

Genevieve Giuliano and Kenneth A. Small. Subcenters in the los angeles region. *Regional Science and Urban Economics*, 21(2):163–182, 1991. ISSN 01660462. DOI: 10.1016/0166-0462(91)90032-I.

Theodore S Glickman. A methodology for estimating time-of-day variations in the size of a population exposed to risk. *Risk Analysis*, 6(3): 317–324, 1986. ISSN 1539-6924.

Hongmian Gong, Cynthia Chen, Evan Bialostozky, and Catherine T Lawson. A gps/gis method for travel mode detection in new york city. *Computers, Environment and Urban Systems*, 36(2):131–139, 2012. ISSN 0198-9715.

Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453 (7196):779–782, 2008. ISSN 0028-0836. DOI: 10.1038/nature06958. URL http://www.nature.com/nature/journal/v453/n7196/full/nature06958.html{%}5Cnhttp://www.nature.com/nature/journal/v453/n7196/pdf/nature06958.pdf.

Stephen Graham. Cities in the real-time age: The paradigm challenge of telecommunications to the conception and planning of urban space. *Environment and Planning A*, 29(1):105–127, 1997. ISSN 0308518X. DOI: 10.1068/a290105.

Stephen Graham and Patsy Healey. Relational concepts of space and place : Issues for planning theory and practice. *European Planning Studies*, 7(5):623–646, 1999. ISSN 0965-4313. DOI: 10.1080/09654319908720542. URL http://www.tandfonline.com/doi/abs/10.1080/09654319908720542.

Stephen Graham and Simon Marvin. *Splintering Urbanism*. Psychology Press, 2001. ISBN 0203452208. DOI: 10.4324/9780203452202. URL http://books.google.com/books?id=6IdAAY9xqlgC{&}pgis=1.

Steve Graham and Simon Marvin. *Telecommunications and the city: Electronic spaces, urban places*. Routledge, 2002. ISBN 1134813937.

Sebastian Grauwin, Stanislav Sobolevsky, Simon Moritz, István Gódor, and Carlo Ratti. Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong. In *Computational Approaches for Urban Environments*, pages 363–387. Springer, 2015. ISBN 9783319114699. DOI: 10.1007/978-3-319-11469-9_15.

Ben Greenstein, Damon McCoy, Jeffrey Pang, Tadayoshi Kohno, Srinivasan Seshan, and David Wetherall. Improving wireless privacy with an identifier-free link layer protocol. In *Proceedings of the 6th international*

conference on Mobile systems, applications, and services, pages 40–53. ACM, 2008. ISBN 1605581399.

Marco Gruteser and Dirk Grunwald. Enhancing location privacy in wireless lan through disposable interface identifiers: a quantitative analysis. *Mobile Networks and Applications*, 10(3):315–325, 2005. ISSN 1383-469X.

Shay Gueron, Simon Johnson, and Jesse Walker. Sha-512/256. In *Proceedings of the 2011 Eighth International Conference on Information Technology: New Generations*, ITNG '11, pages 354–358, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-0-7695-4367-3. DOI: 10.1109/ITNG.2011.69. URL http://dx.doi.org/10.1109/ITNG.2011.69.

Theo Haerder and Andreas Reuter. Principles of transaction-oriented database recovery. *ACM Comput. Surv.*, 15(4):287–317, December 1983. ISSN 0360-0300. DOI: 10.1145/289.291. URL http://doi.acm.org/10.1145/289.291.

Ali Haghani, Masoud Hamedi, Kaveh Sadabadi, Stanley Young, and Philip Tarnoff. Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160(1):60–68, 2010. ISSN 0361-1981.

Jan Erik Håkegård, Tor Andre Myrvoll, and Tor Rune Skoglund. Statistical modelling for estimation of od matrices for public transport using wi-fi and apc data. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1005–1010. IEEE, 2018.

Elaine J Hallisey. Cartographic visualization: an assessment and epistemological review*. *The Professional Geographer*, 57(3):350–364, 2005. ISSN 0033-0124.

Mark Harrower. The cognitive limits of animated maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 42(4):349–357, 2007. ISSN 0317-7173.

Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014. ISSN 1523-0406. DOI: 10.1080/15230406.2014.890072. URL http://www.tandfonline.com/doi/full/10.1080/15230406.2014.890072?src=recsys.

Kevin Hawley and Harold Moellering. A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science*, 32(4):411–423, 2005. ISSN 1523-0406.

S I Hay, A M Noor, A Nelson, and A J Tatem. The accuracy of human population maps for public health application. *Tropical Medicine & International Health*, 10(10):1073–1086, 2005. ISSN 1365-3156.

Tian He, Chengdu Huang, Brian M Blum, John A Stankovic, and Tarek Abdelzaher. Range-free localization schemes for large scale sensor networks. In *Proceedings of the 9th annual international conference on Mobile computing and networking*, pages 81–95. ACM, 2003. ISBN 1581137532.

Janne Heikkilä and Olli Silvén. A real-time system for monitoring of cyclists and pedestrians. *Image and Vision Computing*, 22(7):563–570, 2004. ISSN 0262-8856.

Cesar A Hidalgo and C Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387 (12):3017–3024, 2008. ISSN 0378-4371.

Hande Hong, Girisha Durrel De Silva, and Mun Choon Chan. Crowdprobe: Non-invasive crowd monitoring with Wi-Fi probe. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3): 115, 2018.

Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Guy Pujolle, and Carlo Ratti. Estimating real human trajectories through mobile phone data. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 2, pages 148–153. IEEE, 2013. ISBN 1467360686.

IEEE. IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, Dec 2016.

Md Shahadat Iqbal, Charisma F. Choudhury, Pu Wang, and Marta C. González. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40: 63–74, 2014. ISSN 0968090X. DOI: 10.1016/j.trc.2014.01.002.

Chris Jacobs-Crisioni, Piet Rietveld, Eric Koomen, and Emmanouil Tranos. Evaluating the impact of land-use density and mix on spatiotemporal urban activity patterns: An exploratory study using mobile

phone data. *Environment and Planning A*, 46(11):2769–2785, 2014. ISSN 14723409. DOI: 10.1068/a130309p.

Ralph Jacobson. 2.5 quintillion bytes of data created every day. how does cpg & retail manage it?, Oct 2016. URL https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion.

HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7): 86–94, 2014.

Andreas Janecek, Karin A Hummel, Danilo Valerio, Fabio Ricciato, and Helmut Hlavacs. Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 361–370. ACM, 2012. ISBN 1450312241.

Davy Janssens. *Data Science and Simulation in Transportation Research*. IGI Global, 2013. ISBN 1466649216.

Bing Jia, Baoqi Huang, Hepeng Gao, Wuyungerile Li, and Lifei Hao. Selecting critical wifi aps for indoor localization based on a theoretical error analysis. *IEEE Access*, 2019.

Chen Jia, Yunyan Du, Siying Wang, Tianyang Bai, and Teng Fei. Measuring the vibrancy of urban neighborhoods using mobile phone data with an improved pagerank algorithm. *Transactions in GIS*, 11 2018. DOI: 10.1111/tgis.12515.

Peng Jia, Youliang Qiu, and Andrea E. Gaughan. A fine-scale spatial population distribution on the high-resolution gridded population surface and application in alachua county, florida. *Applied Geography*, 50: 99–107, 2014. ISSN 01436228. DOI: 10.1016/j.apgeog.2014.02.009.

Bin Jiang and Xiaobai Yao. Location-based services and gis in perspective. *Computers, Environment and Urban Systems*, 30(6):712–725, 2006. ISSN 0198-9715.

Shan Jiang, Joseph Ferreira, and Marta C. González. Clustering daily patterns of human activities in the city. In *Data Mining and Knowledge Discovery*, volume 25, pages 478–510. Springer, 2012. ISBN 1384-5810. DOI: 10.1007/s10618-012-0264-z.

Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, page 2. ACM, 2013. ISBN 1450323316.

Shan Jiang, Ana Alves, Filipe Rodrigues, Joseph Ferreira, and Francisco C Pereira. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53:36–46, 2015. ISSN 0198-9715.

Wang Jing, Wang Dianhai, Song Xianmin, and Sun Di. Dynamic od expansion method based on mobile phone location. In *Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on*, volume 1, pages 788–791. IEEE, 2011. ISBN 1612842895.

Warren C. Jochem, Kelly Sims, Edward A. Bright, Marie L. Urban, Amy N. Rose, Phillip R. Coleman, and Budhendra L. Bhaduri. Estimating traveler populations at airport and cruise terminals for population distribution and dynamics. *Natural Hazards*, 68(3):1325–1342, 2013. ISSN 0921030X. DOI: 10.1007/s11069-012-0441-9.

Chaogui Kang, Xiujun Ma, Daoqin Tong, and Yu Liu. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, 391(4):1702–1717, 2012. ISSN 0378-4371.

Nobuo Kawaguchi. Wifi location information system for both indoors and outdoors. In *International Work-Conference on Artificial Neural Networks*, pages 638–645. Springer, 2009.

EK Kim and AM MacEachren. An index for characterizing spatial bursts of movements: A case study with geo-located twitter data. In *GIScience 2014 Workshop on Analysis of Movement Data*, 2014.

Minkyong Kim, David Kotz, Songkuk Kim, Kim Minkyong, David Kotz, and Kim Songkuk. Extracting a mobility model from real user traces. In *INFOCOM*, volume 6, pages 1–13, 2006. ISBN 1424402212. DOI: 10.1109/INFOCOM.2006.173.

Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):2053951714528481, 2014.

Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*, 2010.

Tetsuo Kobayashi, Richard M Medina, and Thomas J Cova. Visualizing diurnal population change in urban areas for emergency management. *The Professional geographer : the journal of the Association of American Geographers*, 63(1):113–30, 2011. ISSN 0033-0124. URL http://www.ncbi.nlm.nih.gov/pubmed/21491706.

Alfred Kobsa. User acceptance of footfall analytics with aggregated and anonymized mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8647 LNCS, pages 168–179, 2014. ISBN 9783319097695. DOI: 10.1007/978-3-319-09770-1_15.

Martijn B W Kobus, Piet Rietveld, and Jos N. Van Ommeren. Ownership versus on-campus use of mobile it devices by university students. *Computers and Education*, 68:29–41, 2013. ISSN 03601315. DOI: 10.1016/j.compedu.2013.04.003.

Constantine E Kontokosta. The quantified community and neighborhood labs: A framework for computational urban planning and civic technology innovation. *SSRN Electronic Journal*, 0(0):1–18, 2015. ISSN 1556-5068. DOI: 10.2139/ssrn.2659896. URL http://papers.ssrn.com/abstract=2659896.

Constantine E Kontokosta and Nicholas Johnson. Urban phenology: Toward a real-time census of the city. *Social Science Research Network*, 2016.

Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009. ISSN 1742-5468.

Johannes Kröckel and Freimut Bodendorf. Visual customer behavior analysis at the point of sale. *International Journal On Advances in Systems and Measurements*, 5(3), 2012.

John Krumm. Inference attacks on location tracks. In *International Conference on Pervasive Computing*, pages 127–143. Springer, 2007.

John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

Tarun Kulshrestha, Divya Saxena, Rajdeep Niyogi, and Jiannong Cao. Real-time crowd monitoring using seamless indoor-outdoor localization. *IEEE Transactions on Mobile Computing*, 2019.

Mei-Po Kwan and Jiyeong Lee. Emergency response after 9/11: the potential of real-time 3d gis for quick emergency response in micro-spatial environments. *Computers, Environment and Urban Systems*, 29(2): 93–113, 2005. ISSN 0198-9715.

Nina Siu-Ngan Lam. Spatial interpolation methods: a review. *The American Cartographer*, 10(2):129–150, 1983. ISSN 0094-1689.

Anthony LaMarca, Yatin Chawathe, Sunny Consolvo, Jeffrey Hightower, Ian Smith, James Scott, Timothy Sohn, James Howard, Jeff Hughes, and Fred Potter. Place lab: Device positioning using radio beacons in the wild. In *International Conference on Pervasive Computing*, pages 116–133. Springer, 2005. ISBN 3540260080.

Renaud Lambiotte, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008. ISSN 03784371. DOI: 10.1016/j.physa.2008.05.014.

Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press, 2014. ISBN 1316094456.

Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010.

Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

Guy Lansley and Paul Longley. Deriving age and gender from forenames for consumer analytics. *Journal of Retailing and Consumer Services*, 30:271–278, 2016a.

Guy Lansley and Paul A Longley. The geography of twitter topics in london. *Computers, Environment and Urban Systems*, 58:85–96, 2016b. ISSN 0198-9715.

Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, and

Others. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, 2012.

Bumsoo Lee. "edge"or "edgless" cities? urban spatial structure in u.s. metropolitan areas, 1980 to 2000*. *Journal of Regional Science*, 47(3): 479–515, 2007. ISSN 00224146. DOI: 10.1111/j.1467-9787.2007.00517.x.

Wouter Lefebvre, B Degrawe, Carolien Beckx, Marlies Vanhulsel, Bruno Kochan, Tom Bellemans, Davy Janssens, Geert Wets, Stijn Janssen, and Ina De Vlieger. Presentation and evaluation of an integrated model chain to respond to traffic-and health-related policy questions. *Environmental modelling & software*, 40:160–170, 2013. ISSN 1364-8152.

Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115:119–133, 2016.

Lin Liao. *Location-based activity recognition*. PhD thesis, University of Washington, 2006.

Cristian Licoppe, Dana Diminescu, Zbignniew Smoreda, and Cezary Ziemlicki. Using mobile phone geolocalisation for 'socio-geographical' analysis of co-ordination, urban mobilities, and social integration patterns. *Tijdschrift voor Economische en Sociale Geografie*, 99(5):584–601, 2008. ISSN 0040747X. DOI: 10.1111/j.1467-9663.2008.00493.x.

Jie Lin and Robert G. Cromley. Evaluating geo-located twitter data as a control layer for areal interpolation of population. *Applied Geography*, 58:41–47, 2015. ISSN 01436228. DOI: 10.1016/j.apgeog.2015.01.006.

Yuming Lin and Weixin Huang. A new perspective of environmental behaviour study: a brief introduction of wi-fi indoor positioning system. *International Journal of Sustainable Society*, 10(4):282–299, 2018.

Liang Liu, Clio Andris, and Carlo Ratti. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6):541–548, 2010. ISSN 0198-9715.

Alyson Lloyd and James Cheshire. Detecting address uncertainty in loyalty card data. *Applied Spatial Analysis and Policy*, Jan 2018. ISSN 1874-4621. DOI: 10.1007/s12061-018-9250-1. URL https://doi.org/10.1007/s12061-018-9250-1.

Amy Lobben. Classification and application of cartographic animation. *The Professional Geographer*, 55(3):318–328, 2003. ISSN 0033-0124.

Sriganesh Lokanathan and Roshanthi Lucas Gunaratne. Mobile network big data for development: Demystifying the uses and challenges. *Digiworld Economic Journal*, 97:75–94, 2015. ISSN 11578637.

Ying Long and Jean-Claude Claude Thill. Combining smart card data and household travel survey to analyze jobs-housing relationships in beijing. *Computers, Environment and Urban Systems*, 53:19–35, 2015. ISSN 01989715. DOI: 10.1016/j.compenvurbsys.2015.02.005.

Paul Longley, James Cheshire, and Alex Singleton. *Consumer Data Research*. UCL Press, 2018.

Paul A Longley, Muhammad Adnan, and Guy Lansley. The geotemporal demographics of twitter usage. *Environment and Planning A*, 47(2):465–484, 2015. ISSN 0308-518X.

Thomas Louail, Maxime Lenormand, Oliva Cantú-García, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *arXiv:1401.4540v1 [physics.soc-ph]*, 18:1–14, 2014. ISSN 2045-2322. DOI: 10.1038/srep05276.

Henghui Lu, Sheng Zhang, Xingchuan Liu, and Xiaokang Lin. Vehicle tracking using particle filter in wi-fi network. In *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd*, pages 1–5. IEEE, 2010. ISBN 1424435730.

Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of GISRUK 2017*. GIS Research UK (GISRUK), 2017.

Alan M MacEachren and Menno-Jan Kraak. Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1): 3–12, 2001. ISSN 1523-0406.

Justin Manweiler, Naveen Santhapuri, Romit Roy Choudhury, and Srihari Nelakuditi. Predicting length of stay at wifi hotspots. In *INFOCOM, 2013 Proceedings IEEE*, pages 3102–3110. IEEE, 2013. ISBN 1467359440.

Stefano Marchetti, Caterina Giusti, Monica Pratesi, Nicola Salvati, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, Luca Pappalardo, and

Lorenzo Gabrielli. Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31(2):263–281, 2015. ISSN 2001-7367.

David Martin. Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, pages 90–97, 1989. ISSN 0020-2754.

David Martin, Samantha Cockings, and Samuel Leung. Developing a flexible framework for spatiotemporal population modeling. *Annals of the Association of American Geographers*, 105(4):754–772, 2015. DOI: 10.1080/00045608.2015.1022089. URL https://doi.org/10.1080/00045608.2015.1022089.

Jeremy Martin, Erik Rye, and Robert Beverly. Decomposition of mac address structure for granular device inference. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pages 78–88. ACM, 2016.

Jeremy Martin, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik C Rye, and Dane Brown. A study of mac address randomization in mobile devices and when it fails. *arXiv preprint arXiv:1703.02874*, 2017.

Doreen Massey. Politics and space/time. *New Left Review*, I(196):249–272, 1992. ISSN 0028-6060. DOI: 10.1049/el:19990302. URL https://newleftreview.org/I/196/doreen-massey-politics-and-space-time.

A Paolo Masucci, Kiril Stanilov, and Michael Batty. Exploring the evolution of london's street network in the information space: A dual approach. *Physical Review E*, 89(1):012805, 2014.

Célestin Matte Mathieu Cunche. On wi-fi tracking and the pitfalls of mac address randomization. In *New Internet Object Challenges: Human-Machine Interaction and Human Factors*, 2016.

Célestin Matte, Mathieu Cunche, Mathy Vanhoef, Franck Rousseau, and Mathy Vanhoef. Defeating mac address randomization through timing attacks. In *ACM WiSec 2016*, pages 15–20. ACM, 2016. ISBN 1450342701.

Owen McCormack. System and method for determining demographic information, March 9 2017. US Patent App. 15/355,652.

John F McDonald. Econometric studies of urban population density: a survey. *Journal of urban economics*, 26(3):361–385, 1989. ISSN 0094-1190.

Malcolm D McIlroy, Elliot N Pinson, and Berkley A Tague. Unix time-sharing system: Foreword. *Bell System Technical Journal*, 57(6):1899–1904, 1978.

Grant McKenzie, Krzysztof Janowicz, Song Gao, and Li Gong. How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54:336–346, 2015. ISSN 0198-9715.

Dermott John James McMeel. The dark web of urban data: Fitness data ecosystems, urban design and privacy in the modern city. *International Journal of Art, Culture and Design Technologies (IJACDT)*, 7(2):12–25, 2018.

Daniel P. McMillen. Nonparametric employment subcenter identification. *Journal of Urban Economics*, 50(3):448–473, 2001. ISSN 0094-1190. DOI: 10.1006/juec.2001.2228. URL http://www.sciencedirect.com/science/article/pii/S0094119001922284.

Daniel P. McMillen. Employment densities, spatial autocorrelation, and subcenters in large metropolitan areas. *Journal of Regional Science*, 44(2): 225–243, 2004. ISSN 00224146. DOI: 10.1111/j.0022-4146.2004.00335.x.

E Mellegard, S Moritz, and M Zahoor. Origin/destination-estimation using cellular network data. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 891–896, 2011. DOI: 10.1109/ICDMW.2011.132.

Jeremy Mennis. Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55(1):31–42, 2003. ISSN 0033-0124.

Jeremy Mennis and Torrin Hultgren. Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, 33(3):179–194, 2006. ISSN 1523-0406.

Mikeazo and Poncho. Formula for the number of expected collisions, 2015. URL https://bit.ly/2YS6zYl.

Harvey J. Miller. The data avalanche is here. shouldn't we be digging? *Journal of Regional Science*, 50(1):181–201, 2010. ISSN 00224146. DOI: 10.1111/j.1467-9787.2009.00641.x.

Prashanth Mohan, Venkata N Padmanabhan, and Ramachandran Ramjee. Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 323–336. ACM, 2008. ISBN 1595939903.

Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646. ACM, 2009. ISBN 1605584959.

David Moore, John Leonard, Daniela Rus, and Seth Teller. Robust distributed network localization with noisy range measurements. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, pages 50–61. ACM, 2004. ISBN 1581138792.

Julie Bauer Morrison, Barbara Tversky, and Mireille Betrancourt. Animation: Does it facilitate learning. In *AAAI spring symposium on smart graphics*, pages 53–59, 2000.

Yaser Mowafi, Ahmad Zmily, Dhiah el Diehn Abou-Tair, and Dirar Abu-Saymeh. Tracking human mobility at mass gathering events using wisp. *Future Generation Communication Technology (FGCT), 2013 Second International Conference on*, pages 157–162, 2013. DOI: 10.1109/FGCT.2013.6767212. URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6767212.

A. B. M. Musa and Jakob Eriksson. Tracking unmodified smartphones using wi-fi monitors. *Proceedings of the 10th ACM Conference on Embedded Networked Sensor Systems (SenSys '12)*, pages 281–294, 2012. DOI: 10.1145/2426656.2426685. URL http://dl.acm.org/citation.cfm?doid=2426656.2426685.

ABM Musa and Jakob Eriksson. Wiflow: real time travel time estimation using Wi-Fi monitors. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pages 429–430. ACM, 2011.

Nicholas N Nagle, Barbara P Buttenfield, Stefan Leyk, and Seth Spielman. Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104(1):80–95, 2014. ISSN 0004-5608.

Fabian Neuhaus. Urbandiary-a tracking project capturing the beat and rhythm of the city: Using gps devices to visualise individual and collective routines within central london. *The Journal of Space Syntax*, 1 (2):336, 2010.

Yibin Ng, Yingchi Pei, and Yunye Jin. Footfall count estimation techniques using mobile data. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 307–314. IEEE, 2017.

Ivan Nikitin, Vitaly Romanov, and Giancarlo Succi. Developing wlan-based intelligent positioning system for presence detection with limited sensors. In *Guide to Ambient Intelligence in the IoT Environment*, pages 95–131. Springer, 2019.

Eamonn O'Neill, Vassilis Kostakos, Tim Kindberg, Alan Penn, Danaë Stanton Fraser, Tim Jones, and Others. Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *International Conference on Ubiquitous Computing*, pages 315–332. Springer, 2006.

J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007a. ISSN 0027-8424.

Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M Argollo De Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179, 2007b. ISSN 1367-2630.

S. Ordonez and A. Erath. Estimating dynamic workplace capacities using public transport smart card data and a household travel survey. *Proceedings of the 17th International Conference of Hong Kong Society for Transportation Studies, HKSTS 2012: Transportation and Logistics Management*, pages 505–512, 2012. ISSN 0361-1981. DOI: 10.3141/2344-03. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84896879480{&}partnerID=tZOtx3y1.

Jeffrey Pang, Ben Greenstein, Ramakrishna Gummadi, Srinivasan Seshan, and David Wetherall. 802.11 user fingerprinting. In *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, pages 99–110. ACM, 2007a. ISBN 1595936815.

Jeffrey Pang, Ben Greenstein, Srinivasan Seshan, and David Wetherall. Tryst: The case for confidential service discovery. In *HotNets*, volume 2, page 1, 2007b.

Luca Pappalardo, Salvatore Rinzivillo, Zehui Qu, Dino Pedreschi, and Fosca Giannotti. Understanding the patterns of car travel. *The European Physical Journal Special Topics*, 215(1):61–73, 2013. ISSN 1951-6355.

Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6:8166, 2015. ISSN 2041-1723. DOI: 10.1038/ncomms9166. URL http://www.nature.com/ncomms/2015/150908/ncomms9166/full/ncomms9166.html{%}5Cnhttp://www.nature.com/doifinder/10.1038/ncomms9166.

R Parrott and F.P Stutz. Urban gis applications. In P. A. Longley, M. F. Goodchild, D.J Maguire, and D.W. Rhind, editors, *Geographic information systems - Principles, techniques, applications and management*, pages 247–60. Wiley, New York, 1999.

Tao Pei, Stanislav Sobolevsky, Carlo Ratti, Shih-Lung Shaw, Ting Li, and Chenghu Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9):1988–2007, 2014. ISSN 1365-8816. DOI: 10.1080/13658816.2014.913794. URL http://www.tandfonline.com/doi/abs/10.1080/13658816.2014.913794.

Santi Phithakkitnukoon and Carlo Ratti. Inferring asymmetry of inhabitant flow using call detail records. *Journal of Advances in Information Technology*, 2(4):239–249, 2011. ISSN 1798-2340.

Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *International Workshop on Human Behavior Understanding*, pages 14–25. Springer, 2010.

Fabio Pinelli, Giusy Di Lorenzo, and Francesco Calabrese. Comparing urban sensing applications using event and network-driven mobile phone location data. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, volume 1, pages 219–226. IEEE, 2015.

Juval Portugali, Han Meyer, Egbert Stolk, and Ekim Tan. *Complexity theories of cities have come of age: an overview with implications to urban planning and design*. Springer Science & Business Media, 2012. ISBN 3642245447.

R Pulselli, P Ramono, Carlo Ratti, and E Tiezzi. Computing urban mobile landscapes through monitoring population density based on

cellphone chatting. *Int. J. of Design and Nature and Ecodynamics*, 3(2): 121–134, 2008.

Rami Puzis, Yaniv Altshuler, Yuval Elovici, Shlomo Bekhor, Yoram Shiftan, and Alex Pentland. Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks. *Journal of Intelligent Transportation Systems*, 17(1):91–105, 2013. ISSN 1547-2450.

Weijun Qin, Jiadi Zhang, Bo Li, and Limin Sun. Discovering human presence activities with smartphones using nonintrusive wi-fi sniffer sensors: The big data prospective. *International Journal of Distributed Sensor Networks*, 2013, 2013. ISSN 15501477. DOI: 10.1155/2013/927940.

John N K Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015. ISBN 1118735722.

Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5): 727–748, 2006. ISSN 02658135. DOI: 10.1068/b32047.

Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLOS ONE*, 5(12):1–6, 12 2010. DOI: 10.1371/journal.pone.0014248. URL https://doi.org/10.1371/journal.pone.0014248.

Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, 2007. ISSN 15361268. DOI: 10.1109/MPRV.2007.53.

Jonathan Reades, Francesco Calabrese, and Carlo Ratti. Eigenplaces: Analysing cities using the space - time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009. ISSN 02658135. DOI: 10.1068/b34133t.

Michael Reibel and Michael E Bufalino. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37(1):127–139, 2005. ISSN 0308-518X.

Jun Rekimoto, Takashi Miyaki, and Takaaki Ishizawa. Lifetag: Wifi-based continuous location logging for life pattern analysis. In *LoCA*, volume 2007, pages 35–49, 2007.

Ian Rose and Matt Welsh. Mapping the urban wireless landscape with argos. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 323–336. ACM, 2010. ISBN 1450303447.

Luca Rossi, James Walker, and Mirco Musolesi. Spatio-temporal techniques for user identification by means of gps mobility data. *EPJ Data Science*, 4(1):11, 2015.

Priya Roy and Chandreyee Chowdhury. Indoor localization for smart-handhelds with stable set of wireless access points. In *2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)*, pages 1–4. IEEE, 2018a.

Priya Roy and Chandreyee Chowdhury. Smartphone based indoor localization using stable access points. In *Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking*, page 17. ACM, 2018b.

Günther Sagl, Bernd Resch, Bartosz Hawelka, and Euro Beinat. From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments. In *Proceedings of the GI-Forum*, pages 54–63, 2012.

Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. Tracking human mobility using wifi signals. *PloS one*, 10(7):e0130824, 2015. ISSN 1932-6203. DOI: 10.1371/journal.pone.0130824. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130824.

T Scott Saponas, Jonathan Lester, Carl Hartung, Sameer Agarwal, and Tadayoshi Kohno. Devices that tell on you: Privacy trends in consumer ubiquitous computing. In *Usenix Security*, volume 3, page 3, 2007.

Saskia Sassen. *The global city: New York, London, Tokyo*, volume 15. Princeton University Press, 2001. ISBN 9780691070636. DOI: 10.2307/2152688. URL http://books.google.com/books?id=PTAiHWK2BYIC.

Marco Luca Sbodio, Francesco Calabrese, Michele Berlingerio, Rahul Nair, and Fabio Pinelli. All aboard: visual exploration of cellphone mobility data to optimise public transport. In *Proceedings of the 19th*

international conference on Intelligent User Interfaces, pages 335–340. ACM, 2014. ISBN 1450321844.

Lorenz Schauer, Martin Werner, and Philipp Marcus. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 171–177, 2014. DOI: 10.4108/icst.mobiquitous.2014.257870. URL http://eudl.eu/doi/10.4108/icst.mobiquitous.2014.257870.

Johannes Schlaich, Thomas Otterstätter, and Markus Friedrich. Generating trajectories from mobile phone data. In *Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies*, 2010.

Robert C Schmitt. Estimating daytime populations. *Journal of the American Institute of Planners*, 22(2):83–85, 1956. ISSN 0002-8991.

Andres Sevtsuk and Carlo Ratti. Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60, 2010. ISSN 1063-0732. DOI: 10.1080/10630731003597322. URL http://www.tandfonline.com/doi/abs/10.1080/10630731003597322.

Kyosuke Shibata and Hiroshi Yamamoto. People crowd density estimation system using deep learning for radio wave sensing of cellular communication. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 143–148. IEEE, 2019.

Masamichi Shimosaka, Takuhiro Kaneko, Kentaro Nishi, Masamichi Shimosaka, Kentaro Nishi, Junichi Sato, and Hirokatsu Kataoka. Extracting land-use patterns using location data from smartphones. *The 1st International Conference on IoT in Urban Space*, IV:1694–1700, 2014. DOI: 10.4108/icst.urb-iot.2014.257220.

Dongyoun Shin, Sofia Georagakopoulou, Daniel Zünd, and Gerhard Schmitt. Crowdsourcing urban sensing: Mobile phone for urban data collection. In *2nd International Hybrid City Conference*, pages 25–28, 2013. ISBN 978-960-99791-1-5.

Seung Hyuck Shin, Chan Gook Park, and Sangon Choi. New map-matching algorithm using virtual track for pedestrian dead reckoning. *ETRI journal*, 32(6):891–900, 2010.

Lei Shu, Yuanfang Chen, Zhiqiang Huo, Neil Bergmann, and Lei Wang. When mobile crowd sensing meets traditional industry. *IEEE Access*, 5: 15300–15307, 2017.

Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012. ISSN 0028-0836.

Alex David Singleton and Paul Longley. The internal structure of greater london: a comparison of national and regional geodemographic models. *Geo: Geography and Environment*, 2(1):69–87, 2015. ISSN 2054-4049.

Alan Smith, David Martin, and Samantha Cockings. Spatio-temporal population modelling for enhanced assessment of urban exposure to flood risk. *Applied Spatial Analysis and Policy*, 9(2):145–163, 2016. ISSN 1874-463X.

Edward Soja. Postmodern geographies, 1989. URL http://books.google.com/books?id=sNcRAQAAIAAJ.

Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6 (10):818–823, 2010a. ISSN 1745-2473. DOI: 10.1038/nphys1760.

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327 (5968):1018–1021, 2010b. ISSN 0036-8075, 1095-9203. DOI: 10.1126/science.1177170.

Balamurugan Soundararaj. Clicker - an Android app for counting pedestrian footfalls with accuracy. https://play.google.com/store/apps/details?id=com.bala.manualcount, 2018.

MS Sruthi. Iot based real time people counting system for smart buildings. *International Journal of Emerging Technology and Innovative Engineering*, 5(2), 2019.

John Steenbruggen, Maria Teresa Borzacchiello, Peter Nijkamp, and Henk Scholten. Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, 78(2):223–243, 2013. ISSN 0343-2521.

John Steenbruggen, Emmanouil Tranos, and Peter Nijkamp. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 39(3-4):335–346, 2015. ISSN 03085961. DOI: 10.1016/j.telpol.2014.04.001.

Shan Suthaharan. Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):70–73, 2014.

Paul Sutton. Modeling population density with night-time satellite imagery and gis. *Computers, Environment and Urban Systems*, 21(3): 227–244, 1997. ISSN 0198-9715.

Paul Sutton, Dar Roberts, C Elvidge, and Kimberly Baugh. Census from heaven: An estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing*, 22(16): 3061–3076, 2001. ISSN 0143-1161.

Karen P Tang, Pedram Keyani, James Fogarty, and Jason I Hong. Putting people in their place: an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 93–102. ACM, 2006.

Ole Tange. *GNU Parallel 2018*. Ole Tange, March 2018. ISBN 9781387509881. DOI: 10.5281/zenodo.1146014. URL https://doi.org/10.5281/zenodo.1146014.

Anna F Tapp. Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science*, 37(3):215–228, 2010. ISSN 1523-0406.

Nicholas Taylor, Richelle Mayshak, Ashlee Curtis, Kerri Coomber, Tanya Chikritzhs, and Peter Miller. Investigating and validating methods of monitoring foot-traffic in night-time entertainment precincts in australia. *International Journal of Drug Policy*, 66:23–29, 2019.

Deborah S K Thomas. Data, data everywhere, but can we really use them. *American hazardscapes: The regionalization of hazards and disasters*, pages 61–76, 2001.

James J Thomas. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005. ISBN 0769523234.

Waldo R Tobler. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530, 1979. ISSN 0162-1459.

Jameson L. Toole, Michael Ulm, Marta C. González, and Dietmar Bauer. Inferring land use from mobile phone activity. In *Proceedings of the*

ACM SIGKDD International Workshop on Urban Computing - UrbComp '12, page 1, 2012. ISBN 9781450315425. DOI: 10.1145/2346496.2346498. URL http://dl.acm.org/citation.cfm?doid=2346496.2346498.

Paul M. Torrens. Wi-fi geographies. *Annals of the Association of American Geographers*, 98(1):59–84, 2008. DOI: 10.1080/00045600701734133.

Anthony M. Townsend. Life in the real-time city: Mobile telephones and urban metabolism. *Journal of Urban Technology*, 7(2):85–104, 2000. ISSN 1063-0732. DOI: 10.1080/713684114.

Emmanouil Tranos. *The geography of the internet: Cities, regions and internet infrastructure in Europe*. Edward Elgar Publishing, 2013.

Emmanouil Tranos and Peter Nijkamp. Urban and regional analysis and the digital revolution: challenges and opportunities. *Hub cities in the knowledge economy: Ashgate, Forthcoming*, 2012.

Emmanouil Tranos and Peter Nijkamp. Mobile phone usage in complex urban systems: a space- time, aggregated human activity study. *Journal of Geographical Systems*, 17(2):157–185, 2015. ISSN 14355949. DOI: 10.1007/s10109-015-0211-9.

Emmanouil Tranos, John Steenbruggen, and Peter Nijkamp. 8 mobile phone operators, their (big) data and urban analysis. *A Research Agenda for Regeneration Economies: Reading City-Regions*, page 140, 2018.

Toivo Vajakas, Tanel Kiis, Amnir Hadachi, and Eero Vainikko. Mobility episode discovery in the mobile networks based on enhanced switching kalman filter. In *2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 1–7. IEEE, 2018.

Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, Frank Piessens, and Piessens Frank. Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pages 413–424. ACM, 2016. ISBN 1450342337. DOI: 10.1145/2897845.2897883.

Leah K VanWey, Ronald R Rindfuss, Myron P Gutmann, Barbara Entwisle, and Deborah L Balk. Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences*, 102(43):15337–15342, 2005. ISSN 0027-8424.

Tien Dang Vo-Huu, Triet Dang Vo-Huu, and Guevara Noubir. Fingerprinting Wi-Fi devices using software defined radios. In *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 3–14. ACM, 2016.

Long Vu, Quang Do, and Klara Nahrstedt. Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*, pages 54–62. IEEE, 2011.

Stephen R. Walli. The posix family of standards. *StandardView*, 3(1): 11–17, March 1995. ISSN 1067-9936. DOI: 10.1145/210308.210315. URL http://doi.acm.org/10.1145/210308.210315.

Neng Wan and Ge Lin. Life-space characterization from cellular telephone collected gps data. *Computers, Environment and Urban Systems*, 39:63–70, 2013. ISSN 0198-9715.

D. Wang, D. Pedreschi, C. Song, Fosca Giannotti, and Albert-László Barabási. Human mobility, social ties, and link prediction. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.*, pages 100–1108, 2011. ISSN 1450308139. DOI: 10.1145/2020408.2020581.

Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.

W. Wang, A.X. Liu, and M. Shahzad. Gait recognition using wifi signals. In *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016. ISBN 9781450344616. DOI: 10.1145/2971648.2971670.

Yan Wang, Jie Yang, Hongbo Liu, and Yingying Chen. Measuring human queues using wifi signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 235–237, 2013. ISBN 9781450319997. DOI: 10.1145/2500423.2504584. URL http://dl.acm.org/citation.cfm?doid=2500423.2504584{%}5Cnhttp://dl.acm.org/citation.cfm?id=2504584.

Amy Wesolowski, Nathan Eagle, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface*, 10 (81):20120986, 2013. ISSN 1742-5689.

Martin Wirz, Tobias Franke, Daniel Roggen, Eve Mitleton-Kelly, Paul Lukowicz, and Gerhard Tröster. Inferring crowd conditions from pedestrians' location traces for real-time crowd monitoring during city-scale mass gatherings. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2012 IEEE 21st International Workshop on*, pages 367–372. IEEE, 2012. ISBN 1467318884.

Luc Johannes Josephus Wismans, LO de Vries, and EC van Berkum. Comparison of interpolation techniques for state estimation on urban networks. In *NECTAR conference 2017*, 2017.

Wei Xi, Yuan He, Yunhao Liu, Jizhong Zhao, Lufeng Mo, Zheng Yang, Jiliang Wang, and Xiangyang Li. Locating sensors in the wild: pursuit of ranging quality. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 295–308. ACM, 2010. ISBN 1450303447.

Zhuliang Xu, Kumbesan Sandrasegaran, Xiaoying Kong, Xinning Zhu, Jingbin Zhao, Bin Hu, and Cheng-Chung Lin. Pedestrain monitoring system using Wi-Fi technology and rssi based localization. *International Journal of Wireless & Mobile Networks*, 5(4):17, 2013.

Yingxiang Yang, David Gerstle, Peter Widhalm, Dietmar Bauer, and Marta Gonzalez. Potential of low-frequency automated vehicle location data for monitoring and control of bus performance. *Transportation Research Record: Journal of the Transportation Research Board*, 2351:54–64, 2013. ISSN 0361-1981.

Yew Yuan, Richard M Smith, and W Fredrick Limp. Remodeling census population with spatial information from landsat tm imagery. *Computers, Environment and Urban Systems*, 21(3):245–258, 1997. ISSN 0198-9715.

Yihong Yuan and Martin Raubal. Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. *International Journal of Geographical Information Science*, 30(8):1594–1621, 2016. ISSN 1365-8816. DOI: 10.1080/13658816.2016.1143555. URL http://www.tandfonline.com/doi/full/10.1080/13658816.2016.1143555.

Sakir Yucel. Measuring benefits, drawbacks and risks of smart community wireless platforms. In *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods (MSV)*, pages 107–113. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2017.

Yang Yue, Han-dong Wang, Bo Hu, Qing-quan Li, Yu-guang Li, and Anthony G O Yeh. Exploratory calibration of a spatial interaction model using taxi gps trajectories. *Computers, Environment and Urban Systems*, 36(2):140–153, 2012. ISSN 0198-9715.

Robbert Zandvliet and Martin Dijst. Short-term dynamics in the use of places: A space-time typology of visitor populations in the netherlands. *Urban Studies*, 43(7):1159–1176, 2006.

Vasileios Zarimpas, Bahram Honary, and Mike Darnell. Indoor 802.1 x based location determination and real-time tracking. In *Wireless, Mobile and Multimedia Networks, 2006 IET International Conference on*, pages 1–4. IET, 2006. ISBN 0863416446.

Zengrzengr, Andy, and Cabral. Calculate distance from rssi, 2017. URL https://bit.ly/2Ohskf9.

Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. Exploring human mobility with multi-source data at extremely large metropolitan scales. *Proceedings of the 20th annual international conference on Mobile computing and networking - MobiCom '14*, pages 201–212, 2014. DOI: 10.1145/2639108.2639116. URL http://dl.acm.org/citation.cfm?id=2639108.2639116.

Kaisa Zhang, Gang Chuai, Weidong Gao, Xuewen Liu, Saidiwaerdi Maimaiti, and Zhiwei Si. A new method for traffic forecasting in urban wireless communication network. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):66, 2019.

Yanting Zhang, Shuaiyu Jin, Yuanyuan Qiao, Kewu Sun, Hao Zhang, and Jie Yang. Exploring urban spatial hotspots' properties using interconnected user-location networks. In *2018 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 84–89. IEEE, 2018.

Huijing Zhao and Ryosuke Shibasaki. A novel system for tracking pedestrians using multiple single-row laser-range scanners. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 35 (2):283–291, 2005. ISSN 1083-4427.

Chen Zhong, Michael Batty, Ed Manley, Jiaqiu Wang, Zijia Wang, Feng Chen, and Gerhard Schmitt. Variability in regularity: Mining temporal mobility patterns in london, singapore and beijing using smart-card data. *PLoS ONE*, 11(2), 2016. ISSN 19326203. DOI: 10.1371/journal.pone.0149222.

Enwei Zhu, Maham Khan, Philipp Kats, Shreya Santosh Bamne, and Stanislav Sobolevsky. Digital urban sensing: A multi-layered approach. *arXiv preprint arXiv:1809.01280*, 2018.